CZECH TECHNICAL UNIVERSITY IN PRAGUE

FACULTY OF TRANSPORTATION SCIENCES

Ing. Petr Šatra

TRAVEL BEHAVIOUR AND APPLICATION OF ITS DETERMINANTS

Doctoral thesis

2020

Declaration

I hereby submit my doctoral thesis, which was made during my studies at CTU in Prague Faculty of Transportation Sciences, to be assessed and defended.

I declare that this thesis has been composed solely by myself and that I have stated all the used bibliographical sources in accordance with Methodical Instructions on Ethical Preparation of University Final Theses.

I have no compelling reason to restrict the usage of this thesis within the meaning of § 60 of Act No. 121/2000 Coll., on Copyright and Rights Related to Copyright and on Amendment to Certain Acts (copyright law).

In Prague on July 31th 2020

…….

signature

Prohlášení

Předkládám tímto k posouzení a obhajobě disertační práci, zpracovanou během studia na ČVUT v Praze Fakultě dopravní.

Prohlašuji, že jsem svou disertační práci vypracoval samostatně a že jsem uvedl veškeré použité informační zdroje v souladu s Metodickým pokynem o etické přípravě vysokoškolských závěrečných prací.

Nemám závažný důvod proti užívání tohoto školního díla ve smyslu § 60 Zákona č.121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon).

V Praze dne 31. července 2020

…….

podpis

CZECH TECHNICAL UNIVERSITY IN PRAGUE

Faculty of Transportation Sciences

TRAVEL BEHAVIOUR
AND APPLICATION OF ITS DETERMINANTS

Doctoral thesis
July 2020
Ing. Petr Šatra

ABSTRACT

The subject of the doctoral thesis **'Travel Behaviour and Application of its Determinants'** is a research of the modal split of the local daily commute to work or school between municipalities of Czechia. The aim is to research the dependencies of modal shares of transport modes on various explanatory variables using decision tree regression and linear regression models. The subject of the thesis is the development of data pre-processing methods (e.g. filtration), which are facilitating the usage of the local commute modal split data, for which is characteristic a large number of origin-destination pairs used by only relatively small average number of travellers. The explanatory variables included 39 socio-demographic indicators and spatial indicators (characteristics of municipalities), and 33 journey characteristic indicators (characteristics of connections between municipalities of origin and destination). The dependencies of the modal share on the characteristics of municipalities of origin (type A models), on the characteristics of municipalities of destination (type B models), and on the characteristics of connections between municipalities (type C models) were researched separately. The dependencies that were found were then applied in a development of recommendations for local governments on a support of transport modes.

KEYWORDS

travel behaviour, modal choice, modal split, modal share, transport planning, transport modelling, origin-destination pair, municipality, region, Population and Housing Census 2011, socio-demographic indicators, spatial indicators, journey characteristic indicators, data pre-processing, decision tree regression, linear regression, dependency modelling

ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

Fakulta dopravní

DOPRAVNÍ CHOVÁNÍ OBYVATEL A APLIKACE JEHO ZÁKONITOSTÍ

disertační práce
červenec 2020
Ing. Petr Šatra

ABSTRAKT

Předmětem disertační práce **'Dopravní chování obyvatel a aplikace jeho zákonitostí'** je výzkum modal splitu lokální denní dojížďky do práce a do školy mezi obcemi ČR. Cílem je nalézt závislosti podílů dopravních módů na dělbě přepravní práce na různých vysvětlujících proměnných pomocí regresních modelů rozhodovacích stromů a lineárních regresních modelů. Předmětem práce je vývoj metod předzpracování dat (např. filtrace), které usnadňují použití dat o lokální dojížďce, pro kterou je charakteristický velký počet dopravních relací používaných pouze relativně malým průměrný počtem cestujících. Vysvětlující proměnné zahrnovaly 39 socio-demografických a prostorových ukazatelů (charakteristiky obcí) a 33 ukazatelů charakteristik cest (charakteristiky spojení mezi obcemi vyjížďky a dojížďky). Závislosti podílů dopravních módů na charakteristikách obcí vyjížďky (modely typu A), na charakteristikách obcí dojížďky (modely typu B) a na charakteristikách spojení mezi obcemi (modely typu C) byly zkoumány samostatně. Nalezené závislosti byly poté aplikovány při návrhu doporučení pro místní samosprávy na podporu jednotlivých druhů dopravy.

KLÍČOVÁ SLOVA

dopravní chování, volba dopravního prostředku, modal split, podíl dopravního prostředku na dělbě přepravní práce, dopravní plánování, dopravní modelování, dopravní relace, obec, kraj, Sčítání lidu, domů a bytů 2011, sociodemografické indikátory, prostorové indikátory, indikátory popisující cestu, před-zpracování dat, regresní rozhodovací stromy, lineární regrese, modelování závislostí

Acknowledgement

Foremost**, I would like to express my sincere gratitude to my life partner Jana Otáhalová, DiS.** for supporting and motivating me in working on my doctoral thesis despite all the things she had to sacrifice to provide me her endless support.

Among many others, I would like to express my sincere thanks to:

my advisor doc. Ing. Jiří Čarský, Ph.D.

for leading my way to publication activity and participation at scientific conferences,

Ms. Petra Nesládková, secretary of the department

for making all the paperwork related to doctoral study travels an enjoyable experience,

Ing. Martin Jacura, Ph.D., head of the department

for his infinite understanding and administrative support of my doctoral studies,

Ms. Leona Šrůtová, librarian

for helpful and enthusiastic assistance with the literature search,

Ing. Mariana Šatrová, Bc. Kateřina Pithartová and Jiří Kapoun

for teaching me how to handle the Python Programming Language,

Ing. Václav Novotný, Ph.D.

for his good advice, which set my doctoral research back on tracks,

Bc. Tomáš 'Jerry' Netrval & Bc. Jan Vavřička

for introducing me to the wonderful world of GIS software ArcMap,

Ing. Sylva Vorlová & Ing. Radek Kuttelwascher

for providing me the right archive version of ArcČR 500 and other GIS data sets,

Ing. Mgr. Pavel Hájek et al.,

for being incredibly helpful in providing me the best data sources from the Czech Statistical Office,

doc. Ing. Josef Kocourek, Ph.D.

for awarding me a scholarship to cover the cost of data acquisition,

Ing. Jaroslav Martínek & Ing. Petr Smítal

for giving me the insights about the bike transport in the Olomouc region,

Ing. Alexandr Vrtěl & Ing. Michal Palán, Ph.D.

for providing me the archive version of the train timetables in machine readable format,

Ing. Jan Buzák & Ing. Vojtěch Hlava

for helpful on demand consultations related to my thesis,

doc. Ing. Ivan Nagy, CSc.

for showing me the power of Bayesian statistics,

Joshua Starmer, Ph.D.

for making StatQuest, the best ever statistics tutorial for dummies like me,

Ing. Bc. Michal Nevoral, M.Sc.

for proving one does not need to be a trainspotter to become a research assistant at our department,

Ing. Filip Müller

for giving hope to all short-sighted visionaries ever dreaming of becoming a pilot,

Ing. Jakub Matějček, M.Sc.

for hearing the calling for help at the utmost moment,

Ing. David Macháček

for sharing his inspiring entrepreneurship spirit,

Ing. Jan Vacek

for being the greatest success of my teaching and leadership during my doctoral studies,

Ing. Ondřej Hába

for advice on the foreword,

Ing. Zdeněk Michl

for waiting for me with finishing his doctoral thesis,

eFDrive, an internal club of the Student union of the CTU in Prague

for providing me an office and cabinet throughout my doctoral studies,

my superiors in my first employer AFRY (AKA good old AF-CITYPLAN), namely:

Ing. Lubomír Tříska, Ing. Monika Bergerová, Ing. Ivo Šimek, CSc. and Bc. Martin Mykisa, DiS.

for creating conditions at our workplace which I have needed for making my thesis to happen,

my superiors in my current employer Deloitte Advisory, namely:

Ing. Jan Janoušek, Ing. Lukáš Filip and Ing. Jakub Sova

for being willing to dismiss me for several weeks when I was finalising the thesis,

my marvellous hosts during that period – my partner's family, namely:

Mrs. Jaroslava Otáhalová, Mr. Igor Otáhal and Mrs. Vladimíra Pavlátová

for taking care of everything while I was just sitting behind the desk and writing the thesis and

Ing. Václav Šíma, their neighbour

for lending me his external monitor to my temporary office.

Last but not the least, **I would like to thank my parents Mgr. Jana Šatrová and Ing. Petr Šatra**, for giving me birth at the first place, for supporting me throughout my life and for limiting their doubts about my doctoral studies to an acceptable amount.

In memory of Ing. Petr Hofhansl, Ph.D. a brilliant mind of transport modelling and inexhaustible source of futuristic visions, which remains my inspiration for ever.

CONTENT

LIST OF ABBREVIATIONS

.csv	comma separated value file format
AKA	also known as
AVG	average
A-S-F	type A model of school commute using filtered set of O-D pairs (A-School-Filtered)
A-W-F	type A model of work commute using filtered set of O-D pairs (A-Work-Filtered)
B-S-F	type B model of school commute using filtered set of O-D pairs (B-School-Filtered)
B-W-F	type B model of work commute using filtered set of O-D pairs (B-Work-Filtered)
CarPass	using the passenger car for commuting as a passenger
CDV	Centrum dopravního výzkumu, v.v.i. (Transport Research Center)
CHMI	Czech Hydrometeorological Institute (Český hydrometeorologický ústav)
CSO	Czech Statistical Office (Český statistický úřad)
C-S-F	type C model of school commute using filtered set of O-D pairs (C-School-Filtered)
C-W-F	type C model of work commute using filtered set of O-D pairs (C-Work-Filtered)
Driver	using the passenger car for commuting as a driver
DTR	decision tree
GDP	gross domestic product
GIS	Geographic Information Systems
IEV	importance of explanatory variable
LR	linear regression
MCT	mass city transport
MEP	municipality with extended powers (obec s rozšířenou působností)
MTB	mountain bike
na	no association
O-D	origin-destination
PS-	aggregation of partial primary and partial secondary sectors of the economy
PT	public transport
PTcom	combination of only public transport modes
PTplus	combinations of public transport modes with private transport modes
r	correlation coefficient
R^2	coefficient of determination
RMSE	root mean square error
SD	standard deviation
ŠTBK	Town of Šternberk or the administrative district of MEP of the same name
T+	aggregation of tertiary, partial quaternary and partial quinary sectors of the economy
Tr	training data set
TT	travel time
TTS	travel time share

[0] Foreword

Three quotes from the book Flights from Olga Tokarczuk were selected to realise the broader perspective of the human travel.

"EVERYWHERE AND NOWHERE

Whenever I set off on any sort of journey I fall off the radar. No one knows where I am. At the point I departed from? Or at the point I'm headed to? Can there be an in-between? Am I like that lost day when you fly east, and that regained night that comes from going west? Am I subject to that much-lauded law of quantum physics that states that a particle may exist in two places at once? Or to a different law that hasn't been demonstrated and that we haven't even thought of yet that says that you can doubly not exist in the same place?" (TOKARCZUK, 2017 p. 48)

"THE RIGHT TIME AND PLACE

Many people believe that there exists in the world's coordinate system a perfect point where time and space reach an agreement. This may even be why these people travel, leaving their homes behind, hoping that even by moving around in a chaotic fashion they will increase their likelihood of happening upon this point. Landing at the right time in the right place – seizing the opportunity, grabbing the moment and not letting go – would mean the code to the safe had been cracked, the combination revealed, the truth exposed. No more being passed by, no more surfing coincidences, accidents and turns of fate. You don't have to do anything – you just have to show up, sign in at that one single configuration of time and place. There you will find your great love, happiness, a winning lottery ticket or the revelation of the mystery everyone's been killing themselves over in vain for all these years, or death. Sometimes in the morning one even has the impression that this moment is close by, that today might be the day it will arrive." (TOKARCZUK, 2017 p. 74)

"ANOTHER COOK

In 1841, Thomas set out on foot to a meeting of the Temperance Society – for he was a great advocate of the temperate mind – from his native Loughborough to Leicester, eleven miles removed. With him went several other gentlemen. Along the way, which was long and tiring, this Cook had an idea – it now seems so strange that no one had ever thought of it before, but that is of course the famous simplicity of brilliant ideas – namely, to rent a railway carriage to transport all the travellers together on the next trip. A month later he managed to ready his first excursion for several hundred people (it is unknown whether all of them were heading to the Temperance Society, however). And so the first travel agency was born." (TOKARCZUK, 2017 p. 224)

[1] Introduction

The **travel behaviour** is a complex phenomenon represented by the user's choice. As the transport users can be influenced, also the travel behaviour is sensitive to outer stimuli such as a construction of a new transport infrastructure or an implementation of new transport system. Such stimuli fall within the scope of a transport development and management. The responsible public governments are managing or should be managing the transport in order to achieve a better mobility for their citizens and visitors. Typically better in terms of effectivity, sustainability, equal opportunity and accessibility and aiming to support the local development.

To do so, the government needs to have a sufficient knowledge, expertise and relevant human resources. However, to build and sustain such a capacity can be a **problem,** as it is not easy nor cheap. The lack of capacity to effectively manage the travel behaviour is especially evident on the lower levels of government, as these are commonly operating with a budget too small to afford the necessary teams of specialists. At the same time, it is the local government who is responsible for the local development, sustainability of rural areas and local transport services accessible on equal terms. It is also trying to reach the highest possible effectivity of the mobility of their citizens and visitors. This research should primarily assist the **local government** to handle this difficult task by providing the necessary knowledge base for an informed decision making. Thus, the **significance of this research** is given by its wide target group of potential beneficiaries represented by all the local governments in Czechia, including 13 self-governing regions and over 6 000 municipalities.

To maximise the impact of this research and the **benefits** for the local government, the research should be focused on the biggest stake of the travel behaviour, which is influenceable by the decision making of the local government. To define the largest portion, let's break down the phenomenon of travel behaviour into four main study fields:

- purpose of the travel,
- frequency of the travel,
- travel time,
- modal split (modal choice).

Usually the biggest decision of the local governments related to the transport policy is how much to spend on the transport in total, but that is rather political issue. The very second decision would be how to distribute the amount among the different modes of transport. Thus, the relevant main stake is closely related to the **modal spit**, which should be the main focus of this research.

The other study fields will also be taken into account. The main purpose of the travel, which represents the majority of all trips, is **travelling to work or school**. The frequency of the travel, which represents the majority of the trips would be the **daily commute**. When it comes to the travel time of the trips, this needs to fall into the category of **local commute** to correlate with the focus on the local government decision making. That narrows down the primary focus of this research to the modal split of the local daily commute to work or school.

In terms of Czechia, the governments of the self-governing regions or simply the '**regions**' are the ones who are primarily responsible for the management of the local transport systems. The regions are managing the infrastructure of the local roads and contracting the public transport in terms of local buses and trains. Basically, what the regions are taking care of is the transport from **municipality to municipality**. The regions are not involved in organization nor in contracting the mass city transport in larger towns and cities, which have their own urban (or even sub-urban) public transport systems. The governments of municipalities (the '**municipalities**') can, on their own discretion, finance or co-finance the local transport systems as an enhancement of the development and operation guaranteed by the regions. That makes the municipalities secondarily responsible for the management of the local transport systems.

The modal split of the journeys between the municipalities can be studied per separate **origin-destination pairs** of municipalities. Such a pair would be a linking of a municipality of origin with a municipality of destination by some transport modes using some specific routings and connections. That makes the origin-destination pair the basic subject of the research. Thus, only the trips to work or school outside the municipality of residence will be studied.

The data describing the travel behaviour within these origin-destination pairs suitable for this research are available from the last national Population and Housing Census in form of traveller flows between the municipalities. For each origin-destination pair, the number of travellers, the mode of transport they have used, and their travel time is available. The **modal share** of transport modes used within these origin-destination pair will be calculated and will represent the **explained variable** or simply the '**Y**' of the research. The ultimate **goal of the research** would be to find and describe the dependencies of modal shares of transport modes on various explanatory variables.

For the **explanatory variable** (the '**X**'), the **characteristics of the municipalities** and **characteristics of the connections between them** will be used. The characteristics related mainly to the socio-demographic, spatial and geographic indicators will be used for the municipalities, while the various traffic engineering, spatial and geographic indicators will be used for the connections. All taken from various reliable and mostly publicly available sources.

How to find the dependencies of the Y's on the X's will be the main **research question** and mapping the ways how to do it will be the main **purpose of the research**. The found dependencies should identify the determinants of the travel behaviour and describe how they are affecting the travel behaviour. It would be then possible to apply the dependencies in predicting the travel behaviour potential in the municipalities and for the connections between them.

Since the input variables will be researched on the level of municipalities and inter-municipal connections, this research approach can be perceived as an attempt to conduct a **macroscopic analysis of travel behaviour**. Using the macroscopic point of view involves less detail than the dominant travel behaviour research using households as the basic subject of research (can be perceived as a microscopic perspective in this thesis). On the other hand, as the macroanalysis of the travel behaviour works with aggregated characteristics of the municipalities and not with the characteristics of specific households, it does not suffer from confidentiality restraints. Thus, the results of the macroanalysis per municipality of any size can be published without the risk of personal data protection breach. Moreover, the population census data allows to process the travel behaviour of much greater portion of the population than in the case of the traditional mobility surveys of households with a limited survey sample.

The possibilities of macroanalysis of travel behaviour shall be explored to verify if the macroanalysis can serve as a complement to existing methods of the travel behaviour research. This also includes a verification of the application of the macroanalysis as a supportive decision-making tool in the transport development and management. The verification shall be done by this **pilot research** of the modal split of local daily commute to work or school between the municipalities of Czechia.

[2] Theoretical background

Firstly, basic concepts will be introduced, and multiple points of view will be presented to stimulate the reflection on the topics. Secondly, analytical methods will be described including the inherent formulas and variables.

2.1 Basic concepts

This chapter guides through the terms and principles, whose understanding is crucial for the subject and purpose of the research.

2.1.1 Travel behaviour

The general definition of the **travel behaviour** was provided by Privitera (2015): "It can generally be referred to as the study of what people do over space and how people use transportation." On the other hand, one might be missing the motivation, rationale or simply the why in the definition. That is what Asad (2013) reminds: "…travel behaviour research usually seeks to find justifications and explanation of people's travel-related options; i.e., how and why rather than how much."

A conceptual overview of the travel behaviour was provided by Axhausen (2007): "Travel behaviour research studies the physical movement of persons outside their reference locations for any purpose. (…) The reference location is the place to which the person returns at the end of the day. This is generally the home (…) The allocation of the time between the first departure of the day from home until the final return between movement and activities defines the outline of a person's daily schedule." This highlight the importance of daily activities, for which the person travels, which is in line with Pickup and Town (1983) who see the travel behaviour as "…an outcome of the balance between the **activity choices** and constraints that face each individual."

The practical outcomes of the travel behaviour have been summarized by Asad (2013) in his review: "People's travel behaviour is typically designated by a number of travel outcomes such as **trip frequency, mode share, journey length** and **time of day** (Meurs & Van Wee, 2003). Travel outcomes such as transport energy consumption and CO_2 emissions have been recently also used as composite metrics (Headicar, Banister, & Pharoah, 2009)."

Unlike many other data utilised in transport sciences, travel behaviour data cannot be collected without direct interaction with travellers. They have to provide the answers to at least the previously stipulated outcomes and perhaps even many more, depending on the type of the survey. Thus,

specialised **mobility surveys** using questionnaires are very demanding type of survey from all perspectives such as methodology, organisation and costs (URBANEK, 2019).

So why is it important to spend significant human and financial resources on conducting mobility surveys and the travel behaviour research in general? The indisputable importance of the travel behaviour data can be seen in the role they play in the creation of transport models (BUZÁK, 2015).

2.1.2 Transport planning and transport modelling

"**Transport planning** is defined as planning required in the operation, provision and management of facilities and services for the modes of transport to achieve safer, faster, comfortable, convenient, economical and environment-friendly movement of people and goods. It is a prediction of usage demand in future travel and to ensure all the necessary facilities and services to cater to that demand. Transport planning is highly essential in shaping cities, enabling economic activities, promoting community interaction, and enhancing quality of life. It is also essential for sustainable development and ensuring safe accessibility at various levels for all individuals. (…) Transportation planning must cover all aspects of city life such as economic development, quality of life, health of public and environment and thereby supporting long-term ecological balance. For this transportation planners and engineers always focus on the efficient movement of people and goods across the country." (TransportPlanning) This empirical definition of transport planning is sufficient introduction into the subject and prepares the ground for the main topic of the chapter, which is the transport modelling.

"What is '**transport modelling**'? Before we dive into transport modelling, let's examine what we mean by 'transport' and what we mean by 'modelling'. **Transport**: Movement of people and of goods between land-uses, e.g. from home to work, from wholesale warehouse to retail outlet, for the purposes of 'gain'; whether that be economic, social, spiritual, recreational, or some other purpose that the person travelling or the organisation consigning the goods considers gainful. It does not cover travel for the purpose of travel itself or for exercise, e.g. a sight-seeing tour bus, a leisure cycle ride, walking the dog, or going jogging. **Modelling**: Any simplified representation of actuality; thus: a photograph is a simplified version of a view; a documentary book is a simplified version of any episode in history, for example; and a house built out of snap-together plastic blocks, no matter how elaborated, is a simplified version of an actual house. A mathematical model is a simplified version of reality expressed numerically and through formulae and algorithms that turn input data into output data that forms information." (HOMER, 2019) To summarize, let's use the definition used by Heyns and Jaarsveld (2017): "Whilst many definitions exist defining transportation modelling it is primarily a mathematical tool using computer software to represent an actual transport system (the real world)

to forecast travel patterns and flows between origins and destinations in geographic space by different modes."

The importance of the **transport models** for transport planning is given by the values of their outcomes: "The outputs from a transport model can provide essential insight into the understanding of an existing or future transport problem, thereby supporting infrastructure design and operational planning. A transport model can also identify the likely impacts that will result from a proposed project, strategy or transport/environmental policy. As such, the transport model plays an essential role as a decision-support tool, providing relevant and accurate information into planning and decision making." (JASPERS, 2014)

For instance, considering the very commonly used **four-step transport forecasting model**, travel behaviour data are required for all the steps. Simply said, following travel behaviour data are required in the respective steps (PTV, 2012):

1) Trip generation – travel behaviour data on number and purpose of trips
2) Trip distribution – travel behaviour data on destinations and purpose of trips
3) Modal choice – travel behaviour data on transport modes used and resulting modal split
4) Route assignment – travel behaviour data on route choice for each trip and mode

In the municipalities and regions, where mobility surveys were conducted, the four-step multimodal transport models can be created, because the travel behaviour data can be used for their calibration. Based on the **travel behaviour data**, distribution curves are developed, to describe the travel behaviour of particular narrow population subgroups, which are showing internal consistency in their composition (demographics) and behaviour (the members of population are travelling from one activity to another, creating so called activity pair) resulting in some specific transport demand. Such travel behaviour patterns typical for some specific population sub-groups and their purpose of travel are referred as demand strata. (BUZÁK, 2015)

The modal choice in the transport models is commonly done based on **generalised costs**, which are aggregating the costs of the travel including the monetary costs such as the fare and the value of travel time. (WARDMAN, et al., 2020) The travel time could be expressed in the form of perceived travel time according to the perception of the real-world travellers, who for example perceive the waiting time on transfers to be longer (and thus more costly) then the time spent in the mean of transport. (BHAT, 1998)

2.1.3 Modal choice

Various definitions of modal choice (mode choice) can be found, describing the relation to the travel behaviour concept. For example, Dave, Raykundaliya and Shah (2013) outlined, that the modal choice is a fundamental element of the travel behaviour. Aston et al. (2019) perceive it in a way that the travel behaviour is a decision-making process and as a such, the choice of a particular transport mode is a subset of the travel behaviour. In regard to the importance of the modal choice research, Milimol, Sreelatha and Soosan (2013) state that: "Modelling traveller's behaviour with respect to mode choice is crucial for effective planning of future transport networks, policy testing and analysis of existing transportation systems." However, modal choice concept can be defined regardless of the superior travel behaviour concept.

An attempt to provide a **complex definition of modal choice** was made in the review from De Witte, Hollevoet, Dobruszkes, Hubert and Macharis (2013): "Based on the specific focus used to analyze the complex process of modal choice in the reviewed articles, three major approaches can be distinguished: a rationalist approach, a socio-geographical approach and a socio-psychological approach. The **rationalist approach** can be seen as the mainstream approach to study modal choice following the assumption that travellers take decisions based on utility maximization attained by minimizing travel time and costs (Shen et al., 2009). This microeconomic approach implies that individuals behave perfectly rational dealing with all kinds of available information (…) and select the alternative with the highest utility. The **socio-geographical approach** explicitly introduces a spatial component into the modal choice decision process and starts from the activity schedule of individuals or households to explain modal choice (Bhat and Singh, 2000; Axhausen, 2002; Meister et al., 2005; Cirillo and Axhausen, 2002). This approach treats the demand for travel as a derived demand, where people are presumed to travel not for the sake of it, but in order to pursue activities distributed in space and time. (…) The **socio-psychological approach** aims at explaining modal choice by the study of attitudes of individuals with regard to the available transport means. The concepts of intentions and habits are important elements in this approach."

Inspired by this multi-disciplinary approach, De Witte et al. (2013) propose to define the modal choice as a decision process to choose between different transport alternatives (various modes and their combinations), which is determined by a combination of interrelated **socio-demographic indicators**, **journey characteristic indicators** and **spatial indicators**, and further influenced by **socio-psychological indicators**. This lays the foundation for design of framework for identification and structuring of the modal choice determinants, comprised of four indicator groups, shown in Figure 1.

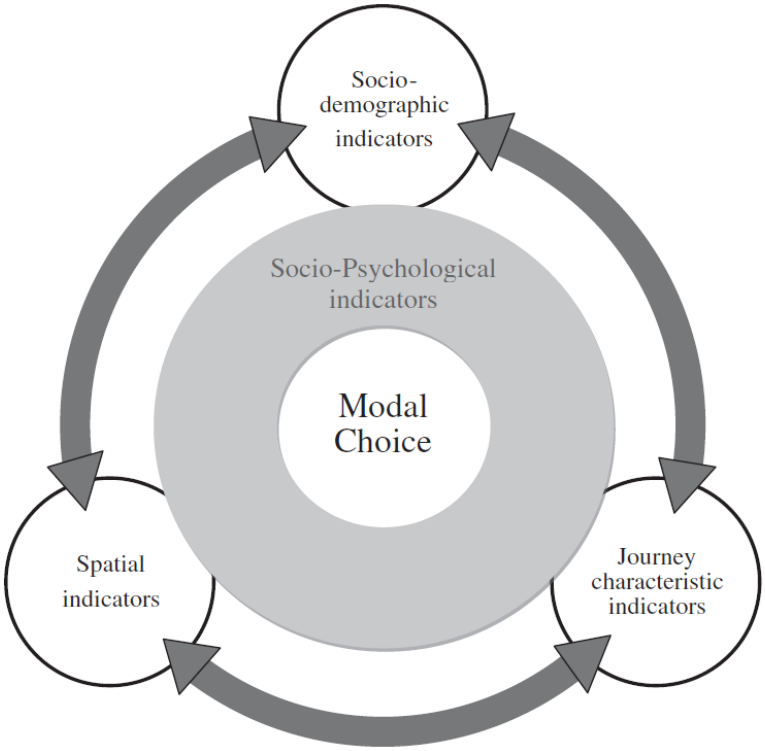


Figure 1: Framework for structuring modal choice determinants by De Witte et al. (2013)

The authors are explaining the graphical representation of the framework as follows: "On the **outside circle**, the framework distinguishes 3 types of determinants constituting the options to make a modal choice. These indicators can be socio-demographic, journey characteristic or space-related. The connection between each of these determinant types represents the possible interrelations and dependencies that may occur among them. (…) The **second circle** represents the influence of socio-psychological or subjective factors, like habits, perceptions, experiences, etc. These factors are determining for how the possibilities shaped in the first circle are acted upon. (…) **Modal choice** is then positioned in the middle as the result of the interactions between the socio-demographic, journey characteristic and spatial indicators combined with the influence of subjective indicators. The aim of this framework is to provide a comprehensible and useable way to structure modal choice determinants to allow for a better understanding of the complex concept of modal choice."

The determinants of modal choice reviewed by De Witte et al. (2013) were then assigned to the indicator groups from the Figure 1 as follows:

<u>Socio-demographic indicators</u>

- **Age** – age of an individual or age groups of people
- **Gender** – gender of an individual or the gender ratio
- **Education** – education of an individual or groups of people with the highest completed level
- **Occupation** – job type, job sector, full-time vs. part time, employment vs. unemployment

- **Income** – income of an individual, household or otherwise defined group of people
- **Household composition** – based on indicators such as age, gender, education or occupation
- **Car availability** – number of cars per driver in household or in area per inhabitant

Spatial indicators

- **Density** – number of inhabitants per defined area (e.g. built-up area, administrative unit etc.)
- **Diversity** – mixed land-use in terms of a diverse pattern of residence, commerce, institutions green space, industry and transport infrastructure accommodated by the neighbourhood
- **Proximity to infrastructure and services** – availability and proximity of road networks and nodes as well as public transport networks and hubs
- **Frequency of public transport** – frequency of the public transport connections
- **Parking** – availability of parking place and facilities and the cost of parking

Journey characteristic indicators

- **Travel motive** – trip purpose based on travel need given by the pursued activity
- **Distance** – trip distance, note that it is strongly related to travel time and travel cost
- **Travel time** – possibly distinguishing the real travel time and the perceived travel time
- **Travel cost** – cost for the use of the travel mode as well as the change in price over time
- **Departure time** – correlation to time of day and the availability of modes at that time
- **Trip chaining** – complex trips with intermediate stops between origin and destination
- **Weather conditions** – such as precipitation and temperature or the different seasons
- **Information** – about the actual availability, reliability and quality of transport alternatives
- **Interchange** – e.g. number, discomfort and waiting time at the interchanges

Socio-psychological indicators

- **Experiences** – past positive or negative experience with the transport mode or provider
- **Familiarity** – the knowledge the user has developed of the various transport modes
- **Lifestyle** – individual's lifestyle
- **Habits** – the strength of the traveller's habit to use the usual mode of transport
- **Perceptions** – quality perception is determined by the person's own history

After providing the framework of modal choice determinants, De Witte et al. (2013) continue with a review of the **influences of determinants** on the modal choice combined with the frequency on which these influences were researched: "In order to get an idea of the importance of these determinants in the modal choice decision, we not only identify the determinants taken into consideration in the different papers, but we also verify in a descriptive way in how many of the

cases each determinant has been labelled as a significant contributor to the modal choice decision." A graphical representation of their classification is in the Figure 2. The authors also add an explanatory example how to read the figure: "Car availability, for instance, is mentioned as a determinant in 36 of the 76 reviewed papers, which corresponds to 47% on the horizontal axis. The vertical axis indicates whether a determinant is frequently recognized as significant. This is expressed as the ratio of the number of times a determinant is found significant against the number of times a determinant is studied. For car availability this ratio stands at 78%, as car availability has been found significant in 28 of the 36 papers where it has been studied."

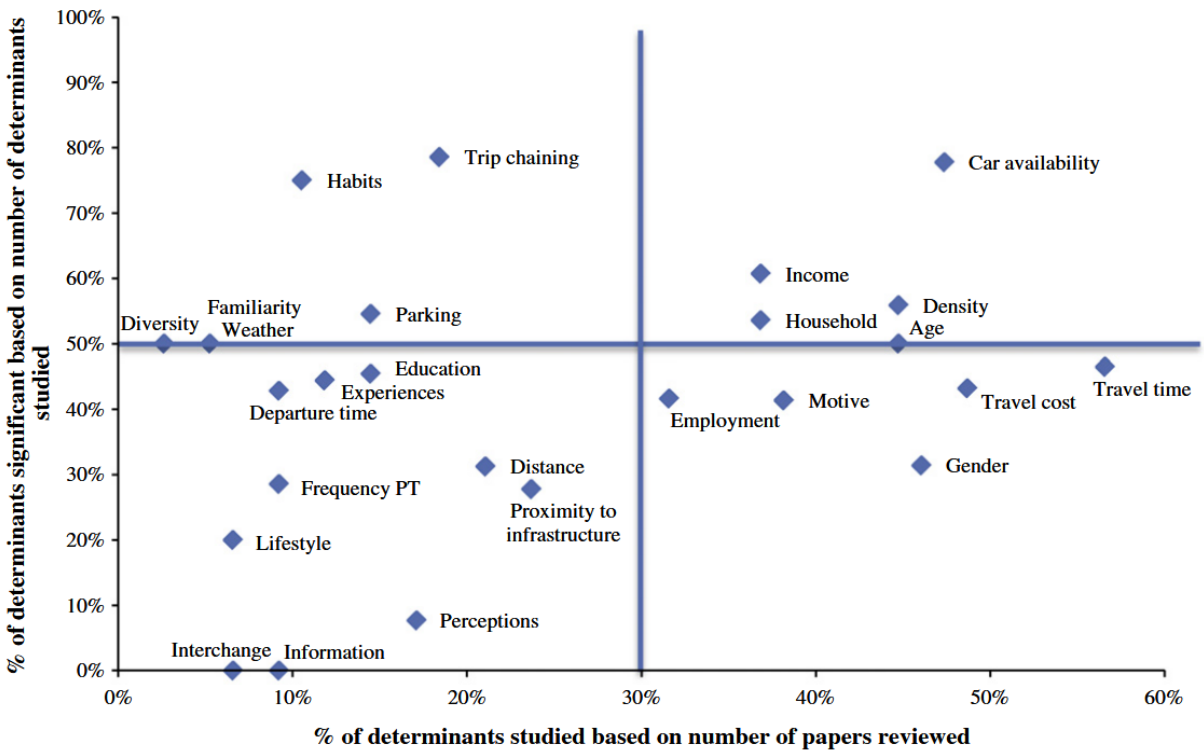


Figure 2: Classification of modal choice determinants based on the review by De Witte et al. (2013)

2.1.4 Modal split

According to Rodrigue, Comtois and Slack (2013), the **modal split** is an outcome of the modal choice. However, the modal choice and modal split should not be treated as synonyms. To further develop this statement, a closer look needs to be taken at **possible definitions** of the modal split.

First definition comes from Vanoutrive (2015), corresponding his workplace perspective to modal split research: "The modal split of a workplace (or city) represents the proportion of individuals using a particular transport mode and is often used as a measure of sustainability..."

Another definition is provided by Matulin, Bošnjak and Simunović (2009) in their study of the modal split of multimodal journeys, where they have stated: “Modal split represents the ratio of different transport modes in the total journey from the origin (O) to the destination (D).”

Rodrigue et al. (2013) have provided three definitions in their broad overview of geography of transport systems. According to them, the modal split could stand for:

“(1) The proportion of total person trips that uses each of various specified modes of transportation.

(2) The process of separating total person trips into the modes of travel used.

(3) A term that describes how many people use alternative forms of transportation.”

The Eurostat glossary connects the definition of the modal split with the shares of different transport modes: “The modal split of transport describes the relative share of each mode of transport for example by road rail or sea.” (2019)

Not forgetting the concept of transport modelling, the modal split is the third step of the four-step transport model, as described by Ungvarai (2019): “In this modelling step a previously calculated amount of trips between an origin and a destination is broken up into parts belonging to different transport modes.”

The provided examples of modal split definitions are mainly showing the **absences of unified approach to modal split definition** in the scientific community, as reported by Matulin et al. (2009) and Ungvarai (2019). This creates a considerable need to clearly define the modal split in every study which involves it, especially considering all the different possible research perspectives or focuses. Without clear definition, the modal split results from different cities, surveys or papers will not be comparable nor easily interpretable and it will not be possible to use it for evaluation of current transport systems and their development, which is one of the main purposes of modal split according to Ungvarai (2019). Framing the modal split figures by the necessary definition is a step to their sound comparison, protecting them from being abused in superficial political statements, for example mentioned by Vanoutrive (2015): “Modal split figures of cities have been used by transport planners and policy-makers to promote their cities as the ‘bicycle capital of the world’.”

To be able to tailor a definition of the modal split for any research purpose, key relevant terms appearing in the above-mentioned definitions need to be defined as well. Firstly, it is necessary to **distinguish** between the **journey** and **trip**. Matulin et al. (2009) put this as follows: “Journey is considered to be a transport of people (including walking) from O-D with the exact purpose. Journey can consist of several trips depending on number of transport modes used…” This is particularly important when multiple modes can be used to travel from the origin to the destination.

Another key part of the framework would be the definition of the **calculation base** of the modal split. It needs to be clearly stipulated, whether the trips made by all mentionable transport modes were included in the calculation or not as reminded by Ungvarai (2019). This is closely related to the **number of modes distinguished in the sample**. Based on one, two, or more modes, Wilken (1977) talks about unimodal, bimodal and multimodal split. It must be noted that no modal choice happens in the case of the **unimodal split**, rather a decision on taking the journey (to travel).

The examples of the most common bimodal and multimodal split are further explained by Reisender Raumplaner (2016) and Ungvarai (2019). In the case of the **bimodal split**, only public transport and private car are usually included in the calculation. Thus, the resulting figure (e.g. "60-40") should be presented as a public transport–private car ratio, rather than a modal split of all trips in the researched sample. It is because the shares of other transport modes (such as walking and cycling) are not included and their omitting is lowering the calculation base, making the result very different from what it should be (BRUNN, 2013).

Even in the case of **multimodal split**, the calculation base must be described. On one hand, the multimodal split includes more modes and thus limits the number of trip omitted from the calculation base, on the other hand, there might be a tendency to omit some unusual transport modes with negligible share from the calculation. In such case, these total shares of omitted modes should be transparently stated, or these residual modes should be aggregated as 'others' and kept in the modal split calculation, making the calculation base equal to the total number of trips (UNGVARAI, 2019).

An essential part of any customized definition of the modal split should be the **number of modes used on a journey** in line with Matulin et al. (2009) and Ungvarai (2019). The above-mentioned 'number of modes distinguished in the sample' does not predetermine the 'number of modes used on a journey'. There could be a multimodal split calculated based on a sample including journeys, which are each consisted of only one trip using single transport mode. The example of that is when a respective methodology takes only one, 'the main', transport mode for each journey but allows for choosing the main one from multiple modes (UNGVARAI, 2019).

However, unless the research focuses on the purely walking trips only, **there is never only one transport mode used for the real-world journey**, simply because every journey always includes some walking, either to a car parking or a public transport stop (UNGVARAI, 2019). Also combining other modes besides walking, e.g. commuter train and tramway in one journey, is nothing extraordinary. Thus, multimodal split allowing for multimodal journeys is often used in the traffic engineering practise. The following explanatory examples are inspired by the work of Matulin et al. (2009).

As stipulated above, each part of a **multimodal journey** completed by a different transport mode is a new trip. In the **example number 1**, a commuter is walking from home to a train station A, then taking a train to a train station B, then walking to a nearby tramway stop 1, then taking a to tramway to tramway stop 2 and then finally walks to work; thus in the end taking five trips and using three transport modes. The walking mode has counted three uses and the remaining two just one. The modal split of such journey is 60 % walking, 20 % train and 20 % tramway. However, the definition of such modal split must be added to provide the necessary context – such modal split is a multimodal split including absolutely all trips taken with different transport modes regardless of their importance to the journey in terms of distance travelled, travel time spent, or cost incurred.

In the **example number 2**, let's add another journey to the sample, connecting exactly the same origin (home) and destination (workplace), but this time composed of three following trips: walking from home to a nearby bus stop α, taking a direct bus to a bus stop β located on the tramway stop 2 and finally walking to work. To make it even more complex, let's say the walking distance from home to bus stop α is shorter than the walking distance from home to the train station A, covered in the previous example 1. The modal split of the second journey is 66.7 % walking and 33.3 % bus, which means, the share of bus on the second journey (33.3 %) is smaller than the sum of the shares of train and tramway (40 %) on the first one, despite the fact the bus has covered a longer distance. Vice versa with the mode walking, which has higher share on second journey despite the fact it has covered a shorter distance than in the case of the first journey.

These discrepancies do not necessarily prove the shortcomings of the modal split concept, but they definitely do point out the necessity of the clearly stated definition of the calculation methodology of the modal split as stated by Matulin et al. (2009) and Ungvarai (2019). This can be best explained by applying different definition to the above-mentioned journey examples. But first, let's calculate the modal split of the sample (first and second journey together) according to the above-stated definition of the modal split (multimodal split **including absolutely all trips** taken with different transport modes). That is going to be based on 5 walking trips, 1 train trip, 1 tramway trip and 1 bus trip, making it in total 8 trips (the calculation base). Share of the mode walking will be 5/8 = 62.5 %, and for the other modes it will be 1/8 = 12.5 %.

Alternative modal split definition could be a multimodal split **including all trips taken with transport modes different from walking**. In such case, the first journey is composed of 2 trips (1 train trip and 1 tramway trip) and the second one is composed of 1 bus trip. The calculation base is 3 trips and the share of each transport mode is 1/3 = 33.3 %. The difference in modal split results based on the two different definitions is evident and so is the possible contradiction these results can make if

presented without the clearly stated definition. Similarly, to the contradictory modal split results of the different studies conducted multiple times in the cities Budapest and Debrecen, provided as an example by Ungvarai (2019).

Since **counting of walking trips** plays a crucial role in the modal split calculation, the authors are usually trying to define some condition or threshold, which would allow to include only significant walking trips into the calculation. A common option is to set a distance threshold. The length of approximately 100 metres is recommended in the European conditions by Ungvarai (2019) as well as by Matulin et al. (2009). The threshold could also be based on the minimum distance generally acceptable for everyone to walk for the particular mode of transport, such as the 300 metres in the case subway stops in Manhattan (OLSZEWSKI, 2007). The other option is to define a condition supporting the importance of the walking trip. Typical example is that **the traveller must make an active decision to use the walking mode** (or any other mode). A short walk from the train to tramway at the public transport transfer hub is not a subject of decision, while a longer walk to train station chosen instead of a short walk to bus stop was selected intentionally based on the traveller's discretion, as suggested Ungvarai (2019). In other words, the walking must be a result of modal choice. Which connects the definition of modal split to the original statement that the modal split is derivative of the modal choice.

It should be noted that the primary goal of any modal split definition is NOT to perfectly reflect the real importance of the transport modes for the whole journey in terms of distance travelled, travel time spent, or cost incurred. The first reason is that with increasing complexity of the travel behaviour data gathered from the respondents, the risk of inconsistent answers increases as well (UNGVARAI, 2019). The second reason is that specialised variables such as passenger-distance [passenger-kilometre] should be used to describe the total performance of the transport modes (UNGVARAI, 2019). As concluded by Ungvarai (2019): "The modal split is definitely measuring behaviour of people and a common perception of travellers. It is connected to local circumstances, mostly to the traffic system of the region. But it is definitely not measuring environmental protection or carbon dioxide emissions, neither distances travelled." **The modal split is measuring the traveller's decisions made in favour of the various transport modes.**

2.1.5 Modal shift

The theoretical backing of the modal shift **(modal change)** was summed up by Rodrigue (2020): "The technological evolution in the transport industry aims at adapting transport infrastructures to growing needs and requirements. When a transport mode becomes more advantageous than another over the same route or market, a modal shift is likely to take place. (…) A modal shift

involves the growth in the demand of a transport mode at the expense of another, although a modal shift can involve an absolute growth in both concerned modes." Which means that even in the case of long term stable modal split, an absolute growth in all concerned modes is taking place if the total demand for transport is growing.

Rodrigue (2020) also covers the **course of the modal shift process**: "A modal shift occurs when one mode has a comparative advantage in a similar market over another. (…) Comparative advantages can involve the difference in cost, time, level of service or reliability between two modes. The higher it is, the more there is an incentive to switch from one mode to the other. Modal shift often takes place over three phases:

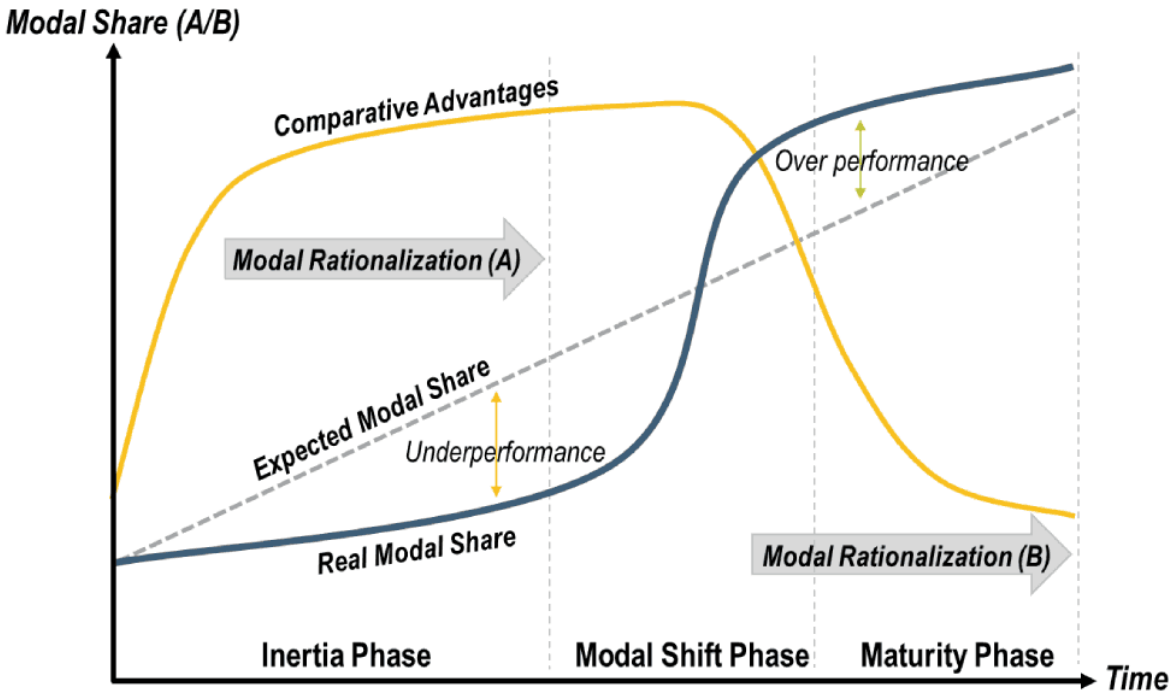


Figure 3: Phases of the modal shift process for the bimodal split according to Rodrigue (2020)

- **Inertia phase.** Initially, a strong level of inertia makes the modal shift a slow and sometimes difficult to perceive process. Only a few users may experiment with modal shift, often as part of a publicly subsidized initiative (e.g. government providing the initial funding to develop services). Inertia implies that the modal shift is often much less significant than expected, leading to a situation of underperformance. The reasons behind inertia are linked to accumulated investments and assets in the existing mode and its terminals. Thus, a corporation will be reluctant to relinquish those assets even if the comparative advantages of the other mode are significant. Management preferences also play a role as expertise was developed to manage flows on the previous mode and may be difficult to adapt to the new mode. (…)

- **Modal shift phase.** This phase represents a fast transition from one mode to the other as the advantages are now acknowledged by the industry. The new transport mode evolves from a situation of underperformance to one of over performance. As inertia involved a modal shift taking place at a rate lower than expected, during the modal shift phase the transition rate is faster than expected. This can take users and authorities by surprise with a rush to cope with additional infrastructure investments. A significant drop in comparative advantages triggers the end of this phase, as the new mode gets increasingly congested and/or as the previous mode loses traffic (closing of some routes, rationalization, price cutting, etc.).
- **Maturity phase.** At this point the market potential is reached with a new equilibrium in modal shares. Their respective comparative advantages are of lesser variance, implying limited incentives to shift cargo or passengers. The focus becomes modal rationalization; using more effectively modal assets."

The modal shift represents the progressive change in modal choice of travellers over the time in a given sample and as such **is influenced by similar factors as the modal choice itself**, as pointed out by Pastori (2018): "The number of factors influencing modal shift and the choice of transport modes is widespread. Key determinants for passenger transport are linked to spatial patterns (e.g. urban density and the proximity to infrastructure and services and journey characteristics) and socio-demographic characteristics (e.g. car ownership, household size, occupation and wage levels)." See chapter 2.1.3 for more determinants of modal choice.

The change of the modal choice, or simply the modal shift, to desirable modal split is a common part of various **transport policies**. A typical goal of contemporary policies is the shift to more sustainable transport systems, as for example mentioned in the work of Kii, Hirota and Minato (2005): "Modal shift from private car to public transport is recognized as one of the key strategy for reduction of energy consumption and CO_2 emission in transport sector." The relief from the negative environmental, social and healthcare impacts of car use are often seen in shifting the modal split from mode car to active travel modes such as walking and cycling as stated by Cools, Moons, Janssens and Wets (2009) and Marshall and Banister (2000). The advantages of such modal shift were summed up in review by Song, Preston and Ogilvie (2017): "A modal shift towards active travel modes such as walking, and cycling has various potential positive impacts. It could reduce air pollution from burning fossil fuels, mitigate traffic congestion, increase levels of physical activity and lead to more sustainable communities (Banister 2008; Rissel 2009; Giles-Corti, Foster, Shilton and Falconer 2010)."

2.2 Analytical methods

The purpose of the use of the presented analytical methods is to search the dependencies of the modal split figures (explained variables or simply the 'Y') and the modal split determinants represented by the explanatory variables (or simply the 'X').

2.2.1 Decision tree regression

"**Decision Trees** are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features." (DecisionTrees) "As the name suggests, we can think of this model as breaking down our data by making a decision based on asking a series of questions." (LI, 2019)

Advantages of the decision trees include:

- The Decision trees are easy to understand and to interpret as they can be visualised in the form of diagrams including easy to read decision rules reached by the model (DecisionTrees)
- A Decision tree model is a kind of a white box model. Any condition of the model can be explained by the Boolean logic. On the contrary, the so-called black box model (employed for example in artificial neural networks) is often much more difficult to understand and to interpret (DecisionTrees)
- The Decision trees are less demanding for data pre-processing as they do not require scaling of data (DHIRAJ, 2019), data normalisation nor any dummy variables to be created (DecisionTrees)
- The cost of using the tree for data prediction is logarithmic in the number of data points used to train the tree (DecisionTrees)

Disadvantages of the decision trees include:

- There is a high risk of overfitting the decision tree model, i.e. creating over-complex trees, which perfectly fit the training data thanks to a complex (overly customized) model structure, but do not represent the general trends in the data (DecisionTrees), see example in Figure 4. The risk of overfitting can be mitigated by pruning the tree, setting the minimum number of samples required at a leaf node or setting the maximum depth of the tree
- Decision tree models are relatively demanding in time needed for the training (DHIRAJ, 2019)
- Instability of the decision tree model, i.e. even a small change in the training data can result in large change of the tree produced (DHIRAJ, 2019)

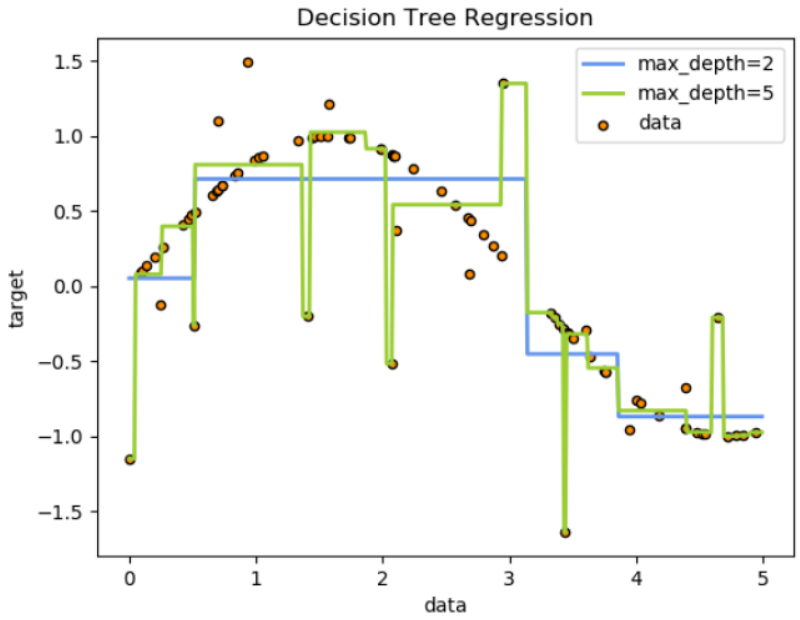


Figure 4: Comparison of two decision tree models learning from an approximate sine curve data. The greater the maximum depth of the tree, the more complex the decision rules and the fitter the model (DecisionTrees)

The visualisation of decision tree model is the **decision tree diagram** (graph), see an example in Figure 5. Following terms are used to describe the elements of the diagram:

- "**Root Node.** It represents entire population or sample, and this further gets divided into two or more homogeneous sets.
- **Splitting.** It is a process of dividing a node into two or more sub-nodes.
- **Decision Node.** When a sub-node splits into further sub-nodes, then it is called decision node.
- **Leaf / Terminal Node.** Nodes do not split is called Leaf or Terminal node (…)
- **Branch / Sub-Tree.** A sub section of entire tree is called branch or sub-tree.
- **Parent and Child Node.** A node, which is divided into sub-nodes is called parent node of sub-nodes whereas sub-nodes are the child of parent node" (BRID, 2018)

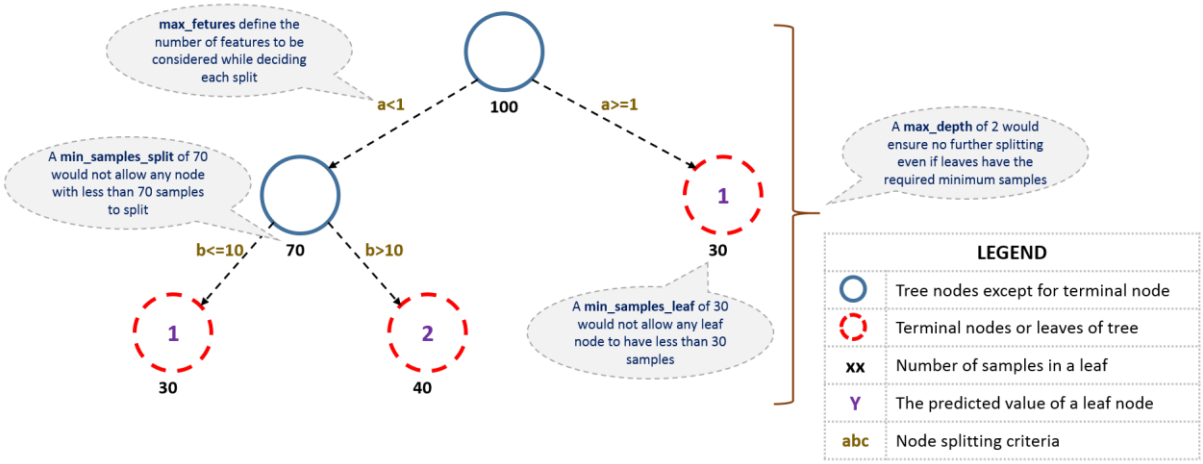


Figure 5: Illustrative example of decision tree diagram by (ANALYTICS VIDHYA, 2016), the top blue tree node is the root node

The decision tree model can use various **decision tree algorithms**, such as ID3, C4.5, C 5.0 or CART. The algorithm, which is typically used for the decision tree regression is CART, which stands for 'Classification and Regression Trees'. The **CART algorithm** supports the numerical target variables used in regression. CART constructs binary trees (parent node has always two child nodes) using the feature and threshold yielding the largest information gain at each node. (DecisionTrees)

Advantage of the CART algorithm include:

- "CART can easily handle both numerical and categorical variables
- CART algorithm identifies the most significant variables and eliminate non-significant ones
- CART can easily handle outliers" (RONAGHAN, 2018)

Disadvantages of the CART algorithm include:

- "CART may have unstable decision tree
- CART splits by one by one variable" (RONAGHAN, 2018)

The **calculation process** of the decision tree model using the CART algorithm can be briefly summarized as follow: "A decision tree is constructed by recursive partitioning — starting from the root node (known as the first parent), each node can be split into left and right child nodes. These nodes can then be further split and they themselves become parent nodes of their resulting children nodes. (…) Starting from the root, the data is split on the feature that results in the largest information Gain G, (explained in more detail below). In an iterative process, we then repeat this splitting procedure at each child node (…) In practice, this can result in a very deep tree with many nodes, which can easily lead to overfitting. Thus, we typically want to prune the tree by setting a limit for the maximal depth of the tree." (LI, 2019)

The **mathematical formulation** of the decision tree regressor model is taken from the webpage of scikit-learn community, which is producing the script and documentation for decision tree modelling in Python programming language:

"Given training vectors $x_i \in R^n$, $i = 1, \dots, l$ and a label vector $y \in R^l$, a decision tree recursively partitions the space such that the samples with the same labels are grouped together.

Let the data at node m be represented by Q. For each candidate split $\theta = (j, t_m)$ consisting of a feature j and threshold t_m, partition the data into $Q_{left}(\theta)$ and $Q_{right}(\theta)$ subsets

$$Q_{left}(\theta) = (x, y) \mid x_j \leq t_m \tag{1}$$

$$Q_{right}(\theta) = Q \setminus Q_{left}(\theta) \tag{2}$$

The impurity at m is computed using an impurity function $H()$, the choice of which depends on the task being solved (classification or regression)

$$G(Q, \theta) = \frac{n_{left}}{N_m} H\left(Q_{left}(\theta)\right) + \frac{n_{right}}{N_m} H\left(Q_{right}(\theta)\right) \tag{3}$$

Select the parameters that minimises the impurity

$$\theta^* = argmin_\theta G(Q, \theta) \tag{4}$$

Recurse for subsets $Q_{left}(\theta^*)$ and $Q_{right}(\theta^*)$ until the maximum allowable depth is reached, $N_m < min_{samples}$ or $N_m = 1$." (DecisionTrees)

The impurity function $H()$ used in the decision tree regression is based on variance reduction using the least squares (Mean Squared Error). "If the target is a continuous value, then for node m, representing a region R_m with N_m observations, common criteria to minimise as for determining locations for future splits is Mean Squared Error, which minimizes the L2 error using mean values at terminal nodes…

$$\bar{y}_m = \frac{1}{N_m} \sum_{i \in N_m} y_i \tag{5}$$

$$H(X_m) = \frac{1}{N_m} \sum_{i \in N_m} (y_i - \bar{y}_m)^2 \tag{6}$$

Where (X_m) is the training data in node m". (DecisionTrees)

One of the key outputs of the decision tree model is the **feature importance** (also referred to as variable importance in line with (VariableImportance)), which is describing the contribution of each

explanatory variable to information gain of the decision tree model. (RONAGHAN, 2018) "Feature importance is calculated as the decrease in node impurity weighted by the probability of reaching that node. The node probability can be calculated by the number of samples that reach the node, divided by the total number of samples. The higher the value the more important the feature." (RONAGHAN, 2018) The first step in calculation of the feature importance is to calculate the node importance (RONAGHAN, 2018):

$$NI_m = w_m H_m - w_{left(m)} H_{left(m)} - w_{right(m)} H_{right(m)} \tag{7}$$

$NI_m =$ the importance of node m

$w_m =$ weighted number of samples reaching node m

$H_m =$ the impurity value of node m

$left_m =$ child node from left split on node m

$right_m =$ child node from right split on node m

a binary tree is assumed (only two child nodes)

The importance of a feature (variable X) is then calculated as according to (RONAGHAN, 2018):

$$FI_j = \frac{\sum_{m:node\ m\ splits\ on\ feature\ j} NI_m}{\sum_{n \in all\ nodes} NI_n} \tag{8}$$

$FI_j =$ the importance of feature j

$NI_m =$ the importance of node m

The feature importance can then be normalised to a value between 0 and 1 if divided by the sum importance values of all features (RONAGHAN, 2018):

$$normFI_j = \frac{FI_j}{\sum_{k \in all\ features} FI_k} \tag{9}$$

2.2.2 Linear regression

"**Linear regression models** are used to show or predict the relationship between two variables or factors. The factor that is being predicted (the factor that the equation solves for) is called the dependent variable. The factors that are used to predict the value of the dependent variable are called the independent variables. In linear regression, each observation consists of two values. One value is for the dependent variable and one value is for the independent variable. In this simple model, a straight line approximates the relationship between the dependent variable and the independent variable." (ANDERSON, et al., 2002) The dependent variable is also being referred to as explained variable and the dependent variable as the explanatory variable. (BONINI, et al., 2018)

The **mathematical formulation** of the simple linear regression model is based on the following formula stated in Bonini, et al. (2018):

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \tag{10}$$

where:

$Y_i =$ dependent variable

$\beta_0 =$ population Y intercept

$\beta_1 =$ population slope coefficient

$X_i =$ independent variable

$\varepsilon_i =$ random error term

A **graphical representation** of the simple linear regression analysis can be seen in Figure 6:

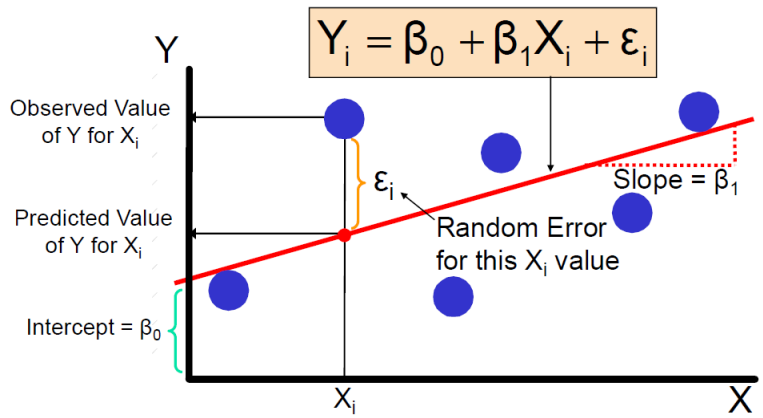


Figure 6: Graphical representation of the simple linear regression analysis (BONINI, et al., 2018)

In this simple linear regression model, a regression line (represented by the red line in Figure 6) is fitted to the data. As stated in Bonini, et al. (2018): "The simple linear regression equation provides an estimate of the population regression line:

$$\hat{Y}_i = b_0 + b_1 X_i \tag{11}$$

where:

$\hat{Y}_i =$ estimated (or predicted) Y value for observation i

$b_0 =$ estimate of the regression intercept

$b_1 =$ estimate of the regression slope

$X_i =$ value of X for observation i"

The estimates of b_0 and b_1 can be found by minimizing the sum of the squared differences between Y and $\hat{Y}$, i.e. using the **least squares method**. The respective formula is provided in Bonini, et al. (2018):

$$min \sum (Y_i - \hat{Y}_i)^2 = min \sum (Y_i - (b_0 - b_1 X_i))^2 \tag{12}$$

The derived formulas for estimates of b_0 and b_1 are then also concluded in Bonini, et al. (2018):

$$b_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \tag{13}$$

$$b_0 = \bar{Y} - b_1 \bar{X} \tag{14}$$

where:

$\bar{X} =$ mean value of the independent variable

$\bar{Y} =$ mean value of the dependent variable

$n =$ number of observations

The **properties of the regression line** were summarized by (StatTrek): "When the regression parameters (b_0 and b_1) are defined as described above, the regression line has the following properties:

- The line minimizes the sum of squared differences between observed values (the Y_i values) and predicted values (the $\hat{Y}_i$ values computed from the regression equation).
- The regression line passes through the mean of the X values ($\bar{X}$) and through the mean of the Y values ($\bar{Y}$).
- The regression constant (b_0) is equal to the y intercept of the regression line.
- The regression coefficient (b_1) is the average change in the dependent variable (Y) for a 1-unit change in the independent variable (X). It is the slope of the regression line."

One of the key outputs of the linear regression is the **Pearson correlation coefficient (r)**, which is indicating the strength of linear correlation between the two variables. It is calculated as a covariance ($cov(X, Y)$) between the variables X and Y 'standardized' by their standard deviations (σ) (SPSStutorials). Thus, the general formulation of the Pearson correlation coefficient is given by the following equation from (RASCHKA, n.d.):

$$r_{X,Y} = \frac{cov(X, Y)}{\sigma_X \sigma_Y} \tag{15}$$

where:

$\sigma_X =$ standard deviation of the variable X

$\sigma_Y =$ standard deviation of the variable Y

After plugging in the formulas for covariance of both variables and their standard deviations, the complex formula for r is then as provided by (SPSStutorials):

$$r_{X,Y} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \tag{16}$$

The linear regression model can show a positive linear relationship, a negative linear relationship, or no **relationship of the two variables** (MIT, 2006). These relationships can be read either from the fitted regression line (MENDENHALL, et al., 2006) or the Pearson correlation coefficient (r) (RASCHKA, n.d.), in these ways:

- **No relationship:**
  - regression line: the fitted line is flat (not sloped)
  - $r = 0$
- **Positive relationship** (value of one variable increases with the increase of the other one):
  - regression line: the line slopes upward
  - $0 < r \leq 1$
- **Negative relationship** (value of one variable increases, while the other one decreases):
  - regression line: the line slopes downward
  - $-1 \leq r < 0$

It is often useful to produce a **correlation matrix** showing the correlations of multiple pairs of variables. However, "Keep in mind that correlations apply to pairs of variables. If you are interested in more than 2 variables, you will probably want to take a look at the correlations between all different variable pairs. These correlations are usually shown in a square table known as a correlation matrix." (SPSStutorials) Example of such correlation matrix can be seen in Figure 7.

Correlation Matrix

	Income 2010	Income 2011	Income 2012	Income 2013	Income 2014
Income 2010	1.000	.913	.857	.770	.720
Income 2011	.913	1.000	.951	.881	.824
Income 2012	.857	.951	1.000	.946	.922
Income 2013	.770	.881	.946	1.000	.968
Income 2014	.720	.824	.922	.968	1.000

Figure 7: Example of a correlation matrix from (SPSStutorials)

"Note that **the diagonal** elements (in red) are the correlations between each variable and itself. Therefore, they are always 1. Also note that the correlations beneath the diagonal (in grey) are redundant because they are identical to the correlations above the diagonal. Technically, we say that

this is a symmetrical matrix. Finally, note that the pattern of correlations makes perfect sense: correlations between yearly incomes become lower insofar as these years lie further apart." (SPSStutorials)

Another key output of linear regression is the **coefficient of determination** (R^2). According to (StatTrek): "It is interpreted as the proportion of the variance in the dependent variable that is predictable from the independent variable.

- The coefficient of determination ranges from 0 to 1.
- An R^2 of 0 means that the dependent variable cannot be predicted from the independent variable.
- An R^2 of 1 means the dependent variable can be predicted without error from the independent variable.
- An R^2 between 0 and 1 indicates the extent to which the dependent variable is predictable. An R^2 of 0.10 means that 10 % of the variance in Y is predictable from X; an R^2 of 0.20 means that 20 % is predictable; and so on.

The formula for computing the coefficient of determination for a linear regression model with one independent variable is given below:

$$R^2 = \left\{ \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \right\}^2 = {r_{X,Y}}^2 \tag{17}$$

If you know the linear correlation (r) between two variables, then the coefficient of determination (R^2) is easily computed using the following formula: $R^2 = r^2$."

StatQuest with Josh Starmer (2019) reminds one of the advantages of using the R^2 over the r, which is that: "R^2 can quantify relationships that are more complicated than simple straight lines." Thus, it can be used for **comparison of strengths** of variable relationships, which were generated by different models.

However, the use of simple linear regression model has some limitation in terms of the following **assumptions** of linear model, reminded by Bonini, et al. (2018):

- **Linearity.** The relationship between X and Y is linear.
- **Independence of errors.** Error values are statistically independent.
- **Normality of error.** Error values are normally distributed for any given value of X.
- **Homoscedasticity of error** (equal variance of error). The probability distribution of the errors has a constant variance.

Chambers and Dinsmore (2014) have summarized the **advantages and disadvantages** of linear regression in the following manner: “The principal advantage of linear regression is its simplicity, interpretability, scientific acceptance, and widespread availability. Linear regression is the first method to use for many problems. (…) Its principal disadvantage is that many real-world phenomena simply do not correspond to the assumptions of a linear model; in these cases, it is difficult or impossible to produce useful results with linear regression.”

[3] Review and evaluation of related research

The search of suitable research sources was conducted using several methods. Besides the well-known Google search engine, two search engines specialised in the scientific research were employed. First of them was Summon research engine, which is a type of 'web scale discovery services' for the academic community. It searches across CTU library catalogues, academic databases such as Web of Science, SCOPUS, ScienceDirect etc., electronic libraries and free access academic databases. The other one was EBSCO search engine, used mainly for searching the literature in the Czech language from local sources.

3.1 Travel behaviour research in Czechia on the national level

The **need for deeper research** of the travel behaviour in Czechia on the national level was already argued by Šenk and Kouřil (2014). They took into consideration the mobility surveys done up to that point in Czechia as well as the travel behaviour research conducted in the neighbouring countries. In they work, they have stated the reasons, why mobility surveys and travel behaviour analysis in sub-areas, such as municipalities and regions, cannot substitute the nationwide survey and analysis. Their reasoning is mainly based on the methodological inconsistency of all those past surveys and researches, procured by various governmental bodies and delivered by multiple vendors, which were not bounded by any common binding methodological guidance. According to their findings, the travel behaviour surveys were done in Czechia mostly on the level of towns, exceptionally cities and only in one case on the regional level. However, different methodologies of these surveys make them nearly impossible to compare. Moreover, any generalisation for the nationwide use is difficult.

The **list of locations** of local mobility surveys done in Czechia nowadays contains several tens of towns and cities, also thanks to the conducted sustainable urban mobility plans in cities like Ostrava (2014), Olomouc (2016), Brno (2017) etc. However, the regional survey remain only the one done in the South Moravian Region in 2013 (CDV, 2020a).

The **first nationwide research** of the travel behaviour in Czechia based on a national mobility survey has been started in 2017 under the leadership of Transport Research Center (CDV), which is a public research institution under the jurisdiction of the Ministry of Transport. This research project was named 'Česko v pohybu' (Czechia in motion). The related mobility surveys were conducted in stages within years 2017 and 2018 – 2019. The first results were published in March 2020 (shortly before submission of this thesis) and results of further analyses should be expected later on (CDV , 2020b).

The mobility surveys under the Czechia in motion project were done according to the **new national 'Methodology of active-travel survey'** prepared by Biler, Kouřil, Rusý, Staněk and Šenk (2014). The new national methodology was delivered as an outcome of the project 'Increasing the efficiency and accuracy of traffic surveys using information and communication technologies', supported by the Technology Agency of the Czech Republic within the public research competition. Thanks to it, the Czechia in motion surveys and results should be comparable with similar travel behaviour research works like Austrian 'KOMOD' and German 'Mobilität in Deutschland' (CDV, 2020a).

The **Czechia in motion surveys work with** responses from **9 419** randomly selected **households** from all over the country. The travel behaviour data were gathered on **22 122 persons** (household members) and **51 434 of their trips**. In order to be publishable, the resulting data were anonymised to assure the personal data protection of the participating households (CDV , 2020b). Thus, the publicly available data and **results are aggregated**, describing the travel behaviour on the country and the regional level only, not on any smaller area.

3.2 Alternative approaches to travel behaviour research

According to Vanoutrive (2015), the "travel behaviour is explained by factors at the trip, personal, household, workplace, neighbourhood and city level." This gives the ground to look at the travel behaviour phenomenon from all these perspectives. Vanoutrive in his work (2015) explores the possibilities, opportunities and limitations of the workplace perspective: "Instead of conceptualising cities as a concentration of residents, the present paper focuses on workplaces and compares the commuting modal split of workplaces that can be found in most cities, such as hospitals banks and supermarkets." In modelling calculations in his paper, he finds some portion of the modal split dependency variations to be explained by the workplace and some also by the cities. This supports the idea of looking at the travel behaviour from the city, or generally the **municipality perspective**.

The tendencies to label the research approaches as **macro and micro analysis** can be found in other research fields. A well related example is the traffic accident analysis undertaken at the macroscopic and microscopic levels, defined for example by Cai (2017): "At the macroscopic level, crashes from a spatial aggregation (such as traffic analysis zone or county) are considered to quantify the impacts of socioeconomic and demographic characteristics, transportation demand and network attributes so as to provide countermeasures from a planning perspective. On the other hand, the microscopic crashes on a segment or intersection are analyzed to identify the influence of geometric design, lighting and traffic flow characteristics with the objective of offering engineering solutions (such as installing sidewalk and bike lane, adding lighting)."

3.3 Managing the mobility

As Santos, Maoh, Potoglou and von Brunn (2013) reminds, "…there is consensus that the private car does not enhance sustainability, whereas public transport and non-motorised modes, such as walking and cycling, do." The sustainability is nowadays an imperative in current strategies for future mobility, such as the sustainable and smart mobility chapter of the European Green Deal (EUROPEAN COMMISSION, 2019). The definition of the **sustainable development** can be drown from Privitera (2015): "The development that meets the needs of the present generations without compromising the ability of future generations to meet their own needs. The definition is the integration of economic efficiency, ecological stability and social equity."

However, **mobility management** does not necessarily mean high-cost investments. Even soft measures such as information campaigns can have positive impact on the travel behaviour of citizens as well as the infrastructural investments (SZAKONYI, et al., 2014).

In regard to the **influence of investments**, Mendiola, González and Cebollada (2014), report that "…municipalities which have access to aboveground / underground rail transport (…) show the highest shares of public transport, thus reducing the proportion of the modal split accounted for by private cars." Which is in line with the findings of Litman and Steele (2013). More examples of stimulation of the travel behaviour by public investments or publicly guaranteed monetary motivation, can be found in chapter 3.6.10.

However, **improper stimulation** of traffic can grow in such dimensions that it massively influences other fields, like land use planning, which then have negative influence back on the travel behaviour. An example of that is discussed in De Vos and Witlox (2013): "Long-distance travel (by car or public transportation) in most western countries is rather cheap due to a number of reasons. First of all, travel costs are low since people do not pay full costs. Car users, for instance, only pay for the internal costs of their travel (e.g., purchase of the car, fuel, insurance), but do not pay the costs they cause to third parties. These external costs, such as congestion, air and noise pollution are paid by all taxpayers and not only by those people who cause them (especially car users). (…) Furthermore, nearly all countries of the world have incorporated social objectives in their public transportation policy (e.g. low transportation rates for labourers living in the countryside), further increasing long-distance travel. Public transportation prices are also often regressive (prices per kilometre are lower for long distances than for short distances), stimulating long-distance travel (e.g. Blauwens et al., 2008; Button, 2010). Since people do not pay full price for car use and public transportation use, an overconsumption of long-distance travel occurs, partly explaining the urban sprawl that has occurred in most western countries in the past decades."

3.4 Defining the major part of the travelling

Travelling to work is generally the most common purpose of the travel. Based on works from Antipova, Wang and Wilmot (2011) and from Habib, Tian and Zaman (2011); Santos et al. (2013) were able to summarize the importance of commute for work as a "major component of daily travel demand and an important source of congestion and pollution", which makes the commute to work desirable for further research.

Looking at the first results of Czechia in motion – the first nationwide mobility survey (CDV , 2020b), the **trips to work have the second highest share** (17.9 %) among all trips, the first being the home trips (40.8 %). Thus, if taking into consideration only the trips for purposes different from 'getting back to home', the share of trips to work would be about 30 %.

As Czechia in motion gathers only trips on daily basis and does not provide any details about how many trips were within one municipality and **how many were inter-municipal**, it is necessary to look for these details in the result of the Population and Housing Census 2011 (CSO , 2013). According to the census, about 85 % of all trips to work or school were made daily and about 55 % of all trips were leaving the municipality of residence. Relying on population census data is actually quite common even abroad, for example Mendiola et al. (2014) were using data from the Spanish Population and Housing Census aggregated at the municipal level.

3.5 Processing the modal split data

The number of travellers commuting from the municipality of origin to the municipality of destination – the **origin-destination flows** – can be in a way considered as some kind of relationship or a tie of those two municipalities. The need for classification of such ties is stipulated in the work of Afonso and Venâncio (2016), allowing (among other things) to define the groups of municipalities with mutually strong commuting ties. They are **assessing the strength of the commuting tie** of each origin-destination pair of municipalities (variable T_{ij}) based on the ratio of number of travellers between municipalities (the origin-destination flow) and the lesser number of inhabitants of the two municipalities. Their assessment is then governed in following way: "As defined in (DORN, 2009), municipalities with stronger ties are the ones with an average value of T_{ij} above 0.02." This method will be presented in more details in chapter 4.1.7.1.

Another method for **assessment of importance of commuting flows** is presented in Mulíček, Seidenglanz, Franke and Malý (2013). They are looking at the five largest flows from each municipality in the South Moravian Region and in the Vysočina Region based on the data from the

Population and Housing Census 2011. These five origin-destination flows are then corelated with all model distributions to find whether the municipality has 1, 2, 3, 4 or 5 significant origin-destination flows. Mulíček et al. (2013) then counted the significant flows from municipalities of the above-mentioned regions to the respective regional capitals. Based on their finding, 388 municipalities out of the 673 of the South Moravian Region have a significant commuting flow to the regional capital, the City of Brno. As for the Vysočina Region, 130 municipalities out of the 704 have a significant commuting flow to its regional capital, the Town of Jihlava.

A proven method of determining the **natural centres** and size **of commuting catchment areas** (commuting hinterlands) is described in the doctoral thesis of Čekal (2006). His work is dedicated to the commuting and migration in the South Bohemian Region. To determine the centre of work commute, he verifies the criterion of 2 500 of occupied jobs in the municipality, first proposed by Řehák (1984) and adds the criterion of positive in-commuting / out-commuting balance and the criterion of at least 10 municipalities in the commuting hinterland of the potential centre. In the end, 16 municipalities in the region have passed these criteria, all being at the same time a **municipality with extended powers** (MEP). Thus, only one MEP out of the total 17 in the South Bohemian Region was not classified as a natural centre of work commute. This shows high correlation of the MEP status and the role of the natural work commute centre. However, the comparison of the extent of administrative districts of these MEPs and their commuting hinterland is not so straightforward, only in one case the administrative districts of MEP includes the same municipalities as its commuting hinterland. The average number of municipalities in the commuting hinterland of the MEP in the South Bohemian Region is 31 (regional capital NOT included among MEPs), while the number of municipalities in commuting hinterland of the regional capital České Budějovice is 120.

The large disproportion between the **size of commuting hinterland** of regional capitals and subordinate centres (such as MEPs) is reflected in the work by De Vos and Witlox (2013). In their study on transportation policy as spatial planning tool, they divides the municipalities into the following categories: “Large cities (cities with more than 100,000 inhabitants); Regional cities (cities with more than 50,000 and less than 100,000 inhabitants); Suburbs (suburban municipalities of big and regional cities); Small cities (cities with less than 50,000 inhabitants) and Countryside.” Then the regional capitals in Czechia with average population size of approx. 133 500 inhabitants (2011, excluding the national capital) fall into the category ‘Large cities’, while the MEPs with average population size of approx. 16 000 inhabitants (2011, excluding the regional capitals, which have dual status) fall into the category ‘Small cities’, different from the regional capitals.

Much larger commuting hinterlands of the regional capitals is an example of extremely large values in the data. The opposite are the extremely small values, offering limited information value. For example, Ruda and Pavlíková (2017) are dealing in their work with basic settlements units with no or extremely small population, which is causing **distortion of the data set**. They have proposed to set a population threshold (100 residents) and exclude the basic settlements units not passing this threshold from their study. They defend this method by the total information value remaining in the data set (99.5 %) after filtering out the sparsely populated basic settlements units.

Last but not least, it is necessary for the modal split research to define the **set of studied transport modes**. When comparing multiple studies, each working with a different set of modes, it is necessary to aggregate the specific transport modes into general categories (e.g. bus and train into public transport) or select only the representative (mostly used) transport modes. Then, the review can consist of only few basic transport modes. For example, Santos et al. (2013) is providing an overview for only 5 key modes (Private car, Public transport, Motorcycle, Bicycle and Foot). On the other hand, the resulting number of transport modes used in the Czechia in motion project (CDV, 2020a) is determined by the number of suggested modes and their allowed combinations in the respective travel behaviour survey. With respect to the frequency of various combinations of transport modes, these main modes were selected for further analysis within the Czechia in motion project: Walking, Bicycle, Mass City Transport, Bus, Train, Car Driver, Car Passenger and Other.

3.6 Explanatory variables (X)

Based on the review of various authors, a large number of explanatory variables was found potentially useful for the modal split research. In order to avoid multiple presentation and discussion of often similar variables, the variables **will be reviewed in groups**, including the suggestions of the authors, who have experience with their use. The occasionally missing influence of the variable on the modal split is given by the fact that some of the variables are taken from land use and local development oriented research works, which are close to the travel behaviour research, but do not study the dependencies of the modal split in particular.

The groups of variables are covering most of the socio-demographic indicators, spatial indicators and journey characteristic indicators and are listed more or less in this order. **No socio-psychological indicators** are included as these are relevant to traveller perspective, but the proposed macroscopic analysis of travel behaviour has the municipality and journey perspective. The groups make a room for the additional variables, recommended by the local development studies.

But first, let's remind the **criteria of suitable transport indicators** (explanatory variables), recommended by Geurs and van Wee (2004): "… first criterion is theoretical backing, namely, a good understanding of what is actually measured is necessary. The second criterion is operationalization, that is, how easily can the indicator be measured in practice. Relevant items are data availability, time and budget. It is clear that a conflict can arise between operational criteria and requirements from theory. Third, an indicator must be understandable and interpretable by researchers, but also by planners and policymakers. Fourth, the impact of trends or policy measures on the phenomenon under study should be readable from the indicator."

3.6.1 Age structure of the population

The usual variable describing the age structure of the population is the **average age**, used for example by De Vos and Witlox (2013). However, this variable does not tell much about the age distribution and thus the extreme age groups are used more commonly as a variable. Typical example is the **share of people with age 65+**, which was used in works by Santos et al. (2013) and Afonsoa and Venâncio (2016). The opposite example is the **share of children aged 0-14**, which was used by Ruda and Pavlíková (2017), Maier and Franke (2015) and Bernard (2012). The way how to combine these two variables together is the **ratio of people with age 65+ and children aged 0-14**, proposed by Novák and Netrdová (2011).

In regard to the influence of the age structure of the population on the modal split, Santos et al. (2013) in the discussion of their results state, that: "…share of public transport is found to decrease with the proportion of elderly population (ELDERLY), in line with Sabir (2011), who finds a negative association between people over 60 and public transport as a mode to travel to work in the Netherlands, and Kim and Ulfarsson (2008), who also find a negative association for people over 65 and public transport for short home-based trips in the Puget Sound region of Washington State in the US." Moreover, Sabir (2011) and Kim and Ulfarsson (2008) both report positive association between the share of people over 60 and 65, respectively and car as well as negative association between the share of elderly and bicycle and walking.

3.6.2 Educational attainment

Just like with the age structure, the inhabitants can be divided into groups according to their highest completed level of education. The completed education level, which most commonly serves as an explanatory variable is represented by the **share of adults who have earned a university degree**. This variable is used by Escalona-Orcaoa et al. (2018), De Vos and Witlox (2013), Maier and Franke (2015), Novák and Netrdová (2011) and Ruda and Pavlíková (2017). The latter ones are also using the **share of adults with only elementary education**.

One can also encounter combined parameters trying to describe the education level of the population of the municipality as a whole, not just one group. Jandová (2016) is using the **average length of the education attendance**. Bernard (2012) suggests the so called **educational index**, which is based on scoring each inhabitant according to his/her highest completed level of education and then summing the scores of all inhabitants in the municipality.

An indirect influence of the age structure can be seen in the relation between the groups of inhabitants with different level of completed education and their availability of car, reported among the results of the Czechia in motion project (CDV , 2020b). The results show higher availability of car in groups with higher level of completed education. The higher car availability might, similarly to higher car ownership, lead to higher share of mode car, see chapter 3.6.7 for more details.

3.6.3 Economic activity of inhabitants

The typical variables describing the economic activity of inhabitants are share of students, share of economically active (working) people, share of unemployed people and share of retired citizens.

The **unemployment rate** (share of unemployed people in the work capable population) is a well-known factor, widely used in many research projects, telling a lot about the state of local economy, demography and job market. It was also used as an explanatory variable in works of De Vos and Witlox (2013), Novák and Netrdová (2011), Jandová (2016) and Bernard (2012).

The **share of students** is used in works by Santos et al. (2013) and De Vos and Witlox (2013). The latter ones are also using the **share of employed** people and the **share of retired** people. The share of retired inhabitants is probably the most commonly used variable, suggested also in Ruda and Pavlíková (2017) and Bernard (2012). A variation on these variables is the **annual average number of working people**, used by Zhikharevich et al. (2015). A special approach is to use the **ratio of economically inactive to active people**, proposed by Jandová (2016).

The share of economically active people (the workers) can be further divided according to the basic sectors. The **share of people working in primary sector** and the **share of people working in quaternary sector** are used in works by Ruda and Pavlíková (2017) and Novák and Netrdová (2011).

The share of economically active people can also be put into relation to the state of the labour market. For instance, Bernard (2012) proposes the **number of occupied jobs per 100 economically active inhabitants**, while Novák and Netrdová (2011) propose the **ratio of number of job vacancies to number of economically active inhabitants.**

To express the entrepreneurship of the municipalities as a whole and of their inhabitants working within the primary sector, Novák and Netrdová (2011) follow the **share of entrepreneurs per 1 000 inhabitants** and the **share of independent farmers per 1 000 inhabitants**.

3.6.4 Average income of inhabitants

Mendiola et al. (2014) are using the **average income of an inhabitant**, Josselin, Rocaboy and Tavéra (2009) worked with the **average income of a voter**, while in the works from Zhikharevich et al. (2015) and from Martarelli and Nagano (2016), the **average wage** is used. A slightly different approach can be seen in the work of De Vos and Witlox (2013) as they decided to use the **number of inhabitants with an income above average**.

Regarding the influence of the income of inhabitants on the modal split, according to the results of Mendiola et al. (2014), the share of private cars is larger in those municipalities where the per capita income is higher.

3.6.5 Gross domestic product

The authors tend to use the **gross domestic product per capita** (GDP per capita) of the municipalities (if available) like in the work of Agovino, Cerciellob and Musella (2019), or GDP per capita of the countries, when comparing the modal split of cities in various countries, like Santos et al. (2013). An alternative to the GDP per capita could be the **volume index of industrial output**, used by Zhikharevich, Rusetskaya and Mladenović (2015).

Regarding the influence of the gross domestic product on the modal split, according to the results of Santos et al. (2013), car share is likely to increase with the GDP per capita.

3.6.6 Tax revenues

Josselin et al. (2009) have used two variables, the **tax base of a median voter** and the **total tax base per inhabitant**, while Bernard (2012) has used the **tax base of self-employed inhabitants.** He highlights the information value of the tax base of self-employed inhabitants over the simple number of self-employed inhabitants in the municipality (see chapter 3.6.3), because besides the information about their number, it also contains the information about their profitability.

3.6.7 Car ownership

A review of use of the **car ownership** as an explanatory variable in the modal split research is provided in the work from Santos et al. (2013).

Regarding the influence of the car ownership on the modal split, Santos et al. (2013) in the discussion of their results state, that: "…city-level share of car is likely to increase as the number of registered cars per 1000 residents (CAR RATE) increases, as expected and in line with Chen et al. (2008), Kain and Liu (2002), Kim and Ulfarsson (2008), Kitamura (2009), Paulley et al. (2006), Pinjari et al. (2007), Sabir (2011), Scheiner (2010), and White (2009)."

3.6.8 Population size

The **number of inhabitants** is used in research works of Santos et al. (2013), Zhikharevich et al. (2015) and Josselin et al. (2009). However, according to Vanoutrive (2015), the population density has proved to be a better predictor then population size.

Regarding the influence of the population size on the modal split, Santos et al. (2013) in the discussion of their results state, that: "The share of public transport (T SHARE) increases with population size, in line Schwanen (2002) and Susilo and Maat (2007)". The latter ones are at the same time reporting negative association of the usage of car with the population size.

3.6.9 Population density

The **gross population density**, calculated from the number of inhabitants and the total area of all municipal grounds is widely used, for example in works of Vanoutrive (2015), Mendiola et al. (2014), De Vos and Witlox (2013) and Gisbert, Martí and Gielen (2017). The latter ones are also using the **net population density** in the same work. Based on the comparison of usage of gross and net population density in one model, Gisbert et al. (2017) consider the net population density to be better performing as an explanatory variable of modal split.

Regarding the influence of the gross population density on the modal split, according to Vanoutrive (2015), the population density is negatively associated with the car use. Mendiola et al. (2014) also report that the share of the commute accounted for by the private cars is larger in municipalities with lower population densities. Moreover, they have stated that: "… the share of public transport in commuting is greater in those municipalities where the population density is higher".

The influence of the net population density on the modal split was summarized in the review from De Witte et al., (2013): "Urban areas (higher density) are usually better served with public transport than rural areas (lower density). One can conclude that public transport is more used in high density areas compared to lower ones (Camagni et al., 2002; Limtanakool et al., 2006)."

3.6.10 Transport services and networks

Variables in this section will generally deal with the accessibility of transport services and networks. In connection with this topic, Vanoutrive (2015) points out: "Measures of accessibility can be seen as a special type of spatial characteristics and give an indication of the ease with which an activity can be reached (Geurs and van Wee, 2004). Accessibility depends on mode, location and time and is considered as a crucial determinant of travel behaviour."

A typical parameter of the accessibility of public transport services in the municipality is the distance to a stop or station of the public transport. The **distance to the closest stop of the public transport** (other than train) and **distance to the closest train stop or station** are variable used by De Vos and Witlox (2013). On the other hand, Vanoutrive (2015) is specifically focusing on the **metro, tram or bus stop within 500 metres from the workplace** and the **railway station within 1km from the workplace**.

A common parameter is the **number of connections by public transport daily**. This variable is used by Santos et al. (2013), Novák and Netrdová (2011) and Mulíček, Seidenglanz, Franke and Malý (2013).

A special parameter of transport services of any kind is their reliability. An attempt to use some reliability metrics can be seen in the work of Vanoutrive (2015). He uses the **presence of congestions** on roads to workplace reported by larger employers in Belgium within their mandatory questionnaire.

The price of the public transport service can be expressed in many ways, depending on the pricing schemes of the operators. Santos et al. (2013) tried to use the **price of monthly ticket for the public transport**.

When it comes to variables describing the length of some transport network, Santos et al. (2013) were looking at the **length of the cycling network**.

Also, the distance to superior transport network could be the explanatory variable. Pecher et al. (2013) are using the **travel time to the closest motorway**, while Vanoutrive (2015) is specifically focusing on the **motorway within 1 km from the workplace**.

A simplified variable describing the access to some transport services in a binary way was shown by Mendiola et al. (2014), who used the variable "…**access to rail transport**, which (…) takes a value of 1 if the municipality has a rail or underground service, and 0 otherwise."

The influences of infrastructure and services on the modal split were reported by Vanoutrive (2015): "As expected, accessibility is a major determinant of mode choice. A railway station in proximity of a workplace decreases car commuting with 5 to 6 per cent, while workplaces located near highways attract 6 per cent more car commuters." De Witte et al. (2013) in their review note: "According to Limtanakool et al. (2006), the availability of a public transport-stop increases public transport use, although for the modal choice decision the proximity to a public transport-stop at the destination side is of greater importance and determining than at the origin side."

The influences on some other variables were reported by Santos et al. (2013), firstly: "…bicycle share (B SHARE) is positively associated with length of the bicycle network (BIKE NETWORK), in line with Pucher et al. (2011)." Secondly: "The share of public transport (T SHARE) … decreases with the cost of a monthly ticket (TRANSIT FARE), in line with Cervero (1998), Chen et al. (2008), Paulley et al. (2006), and Zhang (2004)."

Based on the review of Santos et al. (2013), the public transport service frequency is negatively associated with the share of car and positively associated with share of public transport, which was both reported by Cervero (1998), Kitamura (2009), Paulley et al. (2006) and White (2009).

Last but not least, Mendiola et al. (2014) report that: "…the share of the commute accounted for by private cars is larger in those municipalities with (…) lower availability of public transport…".

3.6.11 Parking

The variable **parking** is usually describing the availability or easiness of finding the parking place in the researched area.

The influence of the availability of parking on the modal split was summarized in the review from De Witte et al., (2013): "The availability of parking has an important impact on modal choice, especially in highly dense areas (Kajita et al., 2004). People are stimulated to use their car (…) when they are guaranteed of having a (free) parking space at work (Kenworthy and Laube, 1996; Kaufmann, 2002; Ye et al., 2007; Vasconcellos, 2005; O'Fallon et al., 2004)."

3.6.12 Real estate

Maier and Franke (2015) in their work suggest using the **share of unoccupied dwellings**. Another variable worth considering is the **property value**, mentioned for example by Mulíček, Seidenglanz, Franke and Malý (2013), Jandová (2016) or Escalona-Orcaoa, Sáez-Pérezb and Sánchez-Valverde García (2018). The importance of **land prices** was discussed in a great depth by De Vos and Witlox (2013): "According to micro-economic location theories, especially developed in the 1960s (see for

example, Alonso, 1964; Wingo, 1961), households trade off transportation costs and land prices in order to find their ideal residential location, whereby living close to a city centre is characterised by high land prices and low transportation costs and living farther away from a city centre with low land prices and high transportation costs."

3.6.13 Geomorphology

These variables are generally describing the terrain and altitude in the municipalities, which could have a major effect on modes such as bike or walking. Pecher, Tasser, Walde and Tappeiner (2013) are using the following variables: **altitude of centre of settlement**, **mean altitude** and **mean slope**.

3.6.14 Climate

The climate in the municipality can be described by long-term averages of climatic variables such as precipitation or temperature. Concerning the precipitation, Santos et al. (2013) are using the **annual average number of rainy days**, while Čarský and Šatra (2019) are using the **annual average precipitation**. The latter ones are also using the **annual average temperature**.

Regarding the influence of the precipitation on the modal split, Santos et al. (2013) have described their experience with annual average number of rainy days as follow: "We did not include RAIN in the specification of bicycle utility because it returned a non-significant parameter. Nonetheless, the evidence regarding the link between rain and cycling is not strong. Sabir (2011) finds a negative association while Pucher et al. (2011) do not report any association."

Čarský and Šatra (2019) report the following influences of climate on modal split: "1) The usage of transport mode Car Driver is decreasing with the decreasing temperature, 2) The usage of mode Car Driver is decreasing with the increasing precipitation, 3) The usage of mode Bus is increasing with the decreasing temperature, 4) The usage of mode Urban Public Transport is decreasing with the increasing temperature."

3.6.15 Travel distance

The variable **travel distance** can be referring to journey or trip distance. The **journey distance** comes in place if the distance between the points of origin and destination is considered. The **trip distance** is relevant to the distance covered by the used transport mode. The trip distance is used when the travel cost is also a subject of research as it is closely related to the distance covered by the used transport mode.

Regarding the influence of the trip distance on the modal split, according to Kim and Ulfarsson (2008), Sabir (2011) and Sheiner (2010), the trip distance is positively associated with the public

transport use and negatively associated with the use of walking. Sabir (2011) and Sheiner (2010) also unanimously report positive association of the trip distance with the private car use. On the other hand, in terms of Bike use, Sabir (2011) reports negative association with the trip distance, while Sheiner (2010) reports positive association up to the distance of 1 500 m and negative association pass this threshold. In terms of thresholds, Rietveld (2000) reports that walking is prevailing up to 1 200 m, bike in the interval 1 200 to 3 700 m and public transport for distance over 3 700 m.

3.6.16 Travel time

The variable **travel time** is closely related to the variable trip distance, based on the average speed of the used transport mode. As Minal and Sekhar (2014) reminds: "… travel time is one of the highly rated factors considered in mode choice and is widely used concept in transportation analysis." As with the modal split, the exact definition of the travel time is necessary, as in-vehicle travel time is used in some papers, while others use total travel time, including for example public transport access and waiting times. According to Bhat (1998), the travellers are more sensitive to the public transport waiting time that the in-vehicle travel time.

3.7 Discussion of the review and recommendation for the methodology

The presented works indicate the **complexity of the travel behaviour** phenomenon, generally too difficult to be described in full depth in the given space. Many of the works depict the influences of multiple variables, out of which many are under direct or indirect control of the local governments. These are the outer stimuli, which the travel behaviour is sensitive to and which can be used for proper management of development and operation of local transport systems.

It should be noted that the regions are not responsible for management of the **local train tracks**, but the level of maintenance of these is more or less connected to the amount of train traffic on them, which is in terms of the prevailing passenger traffic primarily on the decision of the regions, i.e. their order of passenger trains operating under the public service obligation contracts.

As outlined in the introduction, the primary focus of this research should be the modal split of local daily commute to work or school. It is the local governments which are responsible for the transport systems used mainly for the local commute. Thus, the researched data set of origin-destination pairs should be limited to **only pairs relevant to the local commute**. As reported by Čekal (2006), in Czech conditions, the MEPs (municipalities with extended powers) approximately meet the role of natural centres of local commute. Thus, any origin-destination flow, which demonstrably crosses the commuting hinterland of one MEP can be removed from the data set.

It was also stipulated in the introduction, that this research should assist the governments, which are lacking the resources to sustain the necessary teams of specialists on the travel behaviour. Thus, the cities and large towns are not the main priority. Moreover, the size of commuting hinterlands of regional capitals is disproportionally larger than the size of commuting hinterland of an average MEP, as depicted by Čekal (2006). The larger extend of commuting hinterlands of regional capitals prolongs the daily commute trips to regional capitals to a level where these can be considered as long distance trips, with all the consequences discussed by De Vos and Witlox (2013). The **low priority of cities and large towns** and large proportion of long-distance trips among daily commute to regional capitals gives a ground to remove the regional capitals from the data set of origin-destination flows, despite the fact that the regional capitals also have *de jure* the status of MEP. The same, perhaps only on a larger scale, applies to the national capital, the City of Prague.

There are some **large towns in Czechia**, which only have the status of MEP, but about the same population size as smaller regional capitals. Should these also be removed from the data set like the regional capitals? The answer can be given based on the study of Mulíček, Seidenglanz, Franke and Malý (2013) dedicated to Vysočina Region, whose capital Town of Jihlava is the second smallest regional capital in the country with population of only about 50 000 inhabitants. The other large towns in the region are in their population size close to the regional capital, however in the number of the occupied jobs in the town, the Town of Jihlava dominates to the region, thus proving the leading position of the regional capital over the other large towns of the region in terms of the work commute attractivity. Since the other large towns are missing this extra portion of attractivity accounted to only regional capitals, they can be claimed as centres of local importance and stay in the data set.

Last but not least, it is the time to outline the differences in classification of macroscopic and microscopic analysis of the travel behaviour for the needs of this thesis. From the research perspectives outlined by Vanoutrive (2015), the personal and household perspective will be considered as microanalysis, while the workplace (further elaborated by Vanoutrive), the neighbourhood and the **municipal level will be labelled as macroanalysis**. The perspective of trip should be considered independently, as the trips are taken by both members of households (micro units) and members of areas such as workplaces, neighbourhoods and municipalities (macro units). The subject of this research will be the macroanalysis of travel behaviour on journeys between municipalities.

3.7.1 Hypotheses of the influence of explanatory variables on the modal split

The **works devoted to modal split** have suggested 26 variables in 12 groups, including: age structure of the population, educational attainment, average income of inhabitants, gross domestic product, car ownership, population size, population density, transport services and networks, parking, climate, travel distance and travel time.

Moreover, **works devoted to the local development** of municipalities and regions have also been included into the review, since these have the same beneficiary (local governments), detail (municipality) and the ultimate goal of the research, which is to contribute to the local development. These works have enriched the range of suggested variables by another 30 variables in 9 groups: age structure of the population, educational attainment, economic activity of inhabitants, average income of inhabitants, gross domestic product, tax revenues, real estate and geomorphology.

The **suggested variables** are summarized in Table 1. The associations of shares of selected transport modes with explanatory variables suggested by the reviewed works are also provided there. Based on that, associations expected by the author (hypotheses) are stated next to the suggested associations. In the case of available, unambiguous suggestions from the literature, the hypotheses are marked accordingly. In the case of missing suggestions or inconsistent reports of influence, some hypotheses were set at the discretion of the author, based on similarities of the variable or intuitively based on the common sense. The associations, which were not set in accordance with the suggested association, are commented below. It was not possible to set the hypotheses to various combined variables (e.g. Educational index) and variables suggested by the local development studies. Hypotheses were also not set in the case of variables, which were on one hand suggested by the modal split oriented works, on the second hand no data are available on them in the case of this research (more details in chapter 4.2). Thus, a considerable part of the variables remains with no set expectation regarding the influence to modal split.

The expected association (hypothesis) were set at the author's discretion for the following variables, based on the outlined assumption:

- **Property values** and **Land Prices** – as the real estate prices in a municipality have high correlation with its population size (see correlation matrix of municipal related explanatory variables in annex [B]), the expected associations were set accordingly to the suggested associations with the variable population size
- **Total tax base per inhabitant** – this variable is generally expected to be strongly determined by the variable average income of an inhabitant and therefore, its associations were set accordingly to the suggested associations with the average income of an inhabitant

- **Gross population density** – the positive association of the Walking share with this variable was set intuitively
- **Net population density** – the positive association of the Walking share with this variable was set intuitively
- **Unemployment rate** – the positive association of the Public transport share with this variable was set intuitively
- **Share of students** – the positive association of the Public transport share with this variable was set intuitively
- **Share of employed** – the positive association of the Passenger car share with this variable was set intuitively
- **Share of the retired** – as the Share of the retired in a municipality has high correlation with the local share of people with age 65+ (see correlation matrix of municipal related explanatory variables in annex [B]), the expected associations were set accordingly to the suggested associations with the share of people with age 65+
- **Share of entrepreneurs per 1 000 inhabitants** – the positive association of the Passenger car share with this variable was set intuitively

Table 1: Associations of the shares of selected transport modes with the explanatory variables suggested by the reviewed works

Associations of suggested explanatory variables with shares of selected transport modes		**Passenger car**		**Public transport**		**Bicycle**		**Walking**	
Group	**Variable**	**Suggested assoc.**	**Expected assoc.**	**Suggested assoc.**	**Expected assoc.**	**Suggested assoc.**	**Expected assoc.**	**Suggested assoc.**	**Expected assoc.**
Age structure	Average age	na	na	na	na	na	na	na	na
Age structure	Share of people with age 65+	+	+	-	-	-	-	-	-
Age structure	Share of children aged 0-14	na	na	na	na	na	na	na	na
Age structure	Ratio of people with age 65+ and children aged 0-14	na	na	na	na	na	na	na	na
Education	Share of adults who have earned a university degree	+	+	na	na	na	na	na	na
Education	Share of adults with only elementary education	-	-	na	na	na	na	na	na
Education	Average length of the education attendance	na	na	na	na	na	na	na	na
Education	Educational index	na	na	na	na	na	na	na	na
Economic activity	Unemployment rate	na	na	na	+	na	na	na	na
Economic activity	Share of students	na	na	na	+	na	na	na	na
Economic activity	Share of employed	na	+	na	na	na	na	na	na
Economic activity	Share of the retired	na	+	na	-	na	-	na	-
Economic activity	Annual average number of working people	na	na	na	na	na	na	na	na
Economic activity	Ratio of economically inactive to active people	na	na	na	na	na	na	na	na
Economic activity	Share of people working in primary sector	na	na	na	na	na	na	na	na
Economic activity	Share of people working in quaternary sector	na	na	na	na	na	na	na	na
Economic activity	Number of occupied jobs per 100 economically active inhabitants	na	na	na	na	na	na	na	na
Economic activity	Ratio of number of job vacancies to number of economically active inhabitants	na	na	na	na	na	na	na	na
Economic activity	Share of entrepreneurs per 1 000 inhabitants	na	+	na	na	na	na	na	na

Associations of suggested explanatory variables with shares of selected transport modes		Passenger car		Public transport		Bicycle		Walking	
Group	**Variable**	**Suggested assoc.**	**Expected assoc.**	**Suggested assoc.**	**Expected assoc.**	**Suggested assoc.**	**Expected assoc.**	**Suggested assoc.**	**Expected assoc.**
Economic activity	Share of independent farmers per 1 000 inhabitants	na	na	na	na	na	na	na	na
Income of inhab.	Average income of an inhabitant	+	na	~	na	-	na	+	na
Income of inhab.	Average income of a voter	na	na	na	na	na	na	na	na
Income of inhab.	Average wage	na	na	na	na	na	na	na	na
Income of inhab.	Number of inhabitants with an income above average	na	na	na	na	na	na	na	na
GDP	Gross domestic product per capita	+	na	~	na	na	na	na	na
GDP	Volume index of industrial output	na	na	na	na	na	na	na	na
Tax revenues	Tax base of a median voter	na	na	na	na	na	na	na	na
Tax revenues	Total tax base per inhabitant	na	+	na	na	na	-	na	+
Tax revenues	Tax base of self-employed inhabitants	na	na	na	na	na	na	na	na
Car ownership	Car ownership	+	+	-	-	~	na	~	na
Population size	Number of inhabitants	-	-	+	+	~	na	~	na
Population density	Gross population density	-	-	+	+	na	na	na	+
Population density	Net population density	na	na	+	+	na	na	na	+
Transport services	Distance to the closest stop of the public transport	na	na	-	-	na	na	na	na
Transport services	Distance to the closest train stop or station	na	na	-	-	na	na	na	na
Transport services	Metro, tram or bus stop within 500 metres from the workplace	na	na	na	na	na	na	na	na
Transport services	Railway station within 1km from the workplace	-	na	+	na	na	na	na	na
Transport services	Number of connections by public transport daily	-	-	+	+	na	na	na	na
Transport services	Presence of congestions	na	na	na	na	na	na	na	na

Associations of suggested explanatory variables with shares of selected transport modes		Passenger car		Public transport		Bicycle		Walking	
Group	**Variable**	**Suggested assoc.**	**Expected assoc.**	**Suggested assoc.**	**Expected assoc.**	**Suggested assoc.**	**Expected assoc.**	**Suggested assoc.**	**Expected assoc.**
Transport services	Price of monthly ticket for the public transport	+	+	-	-	na	na	na	na
Transport services	Length of the cycling network	na	na	na	na	+	+	na	na
Transport services	Travel time to closest motorway	na	na	na	na	na	na	na	na
Transport services	Motorway within 1 km from the workplace	+	na	na	na	na	na	na	na
Transport services	Access to rail transport	+	na	na	na	na	na	na	na
Parking	Availability of parking place and facilities	+	na	-	na	na	na	na	na
Real estate	Share of unoccupied dwellings	na	na	na	na	na	na	na	na
Real estate	Property values	na	na	na	na	na	na	na	na
Real estate	Land prices	na	na	na	na	na	na	na	na
Geomorphology	Altitude of centre of settlement	na	na	na	na	na	na	na	na
Geomorphology	Mean altitude	na	na	na	na	na	na	na	na
Geomorphology	Mean slope	na	na	na	na	na	na	na	na
Climate	Annual average number of rainy days	na	na	na	na	~	na	na	na
Climate	Annual average precipitation	-	-	na	na	na	na	na	na
Climate	Annual average temperature	+	+	-	-	na	na	na	na
Travel distance	Trip distance	+	+	-	-	~	na	-	-
Travel time	Total travel time	na	na	na	na	na	na	na	na

Variables suggested by modal split oriented research works
Variables suggested by local development oriented research works

+	positive association
-	negative association
~	inconsistent reports of influence
na	no association available

[4] Research approach and methodology

As discussed earlier, the research of the travel behaviour will be mainly focused on the research of dependencies of the modal split on the selected explanatory variables. Finding and describing such dependencies could lay foundation for creation of a decision-making tool for transport policies adopted by the self-governing units such as regions and municipalities. This chapter present the **procedures and methods developed or selected by the author** of the thesis for the purpose of finding the dependencies, including the data pre-processing methods, which were a key prerequisite for the dependency research.

4.1 Collection and pre-processing of explained variable data set

The modal split data will be set as explained variable and as such, its dependency on the explanatory variables will be modelled. But first, the explained variable data set needs to be pre-processed.

4.1.1 Data gathering and description

The raw data on travel behaviour of all Czech citizens were gathered in March 2011 during the last national **Population and Housing Census 2011** organized by the Czech Statistical Office. The decisive moment was the midnight from March 25 to March 26. The questions related to travel behaviour (number 21 through 24) were included in the so-called **Census Form – Persons** (CSO , 2011a). The example of such is included in the annex [A] along with its translation into English.

The questions in the form were introduced as follows: "Questions 21, 22, 23 and 24 about commuters to work or school are to be filled out only by employees and pupils, students and apprentices. Working students and apprentices are to answer according to the commute to school."

The **question 21** was dedicated to finding the address of **Location of place of work or school**. There was also option to state a 'job without permanent workplace'.

The questions 22 and 23 were specifically introduced as follows: "In regard to daily commuting to work or school, state:" The **question 22** on the modal split had a following wording: "**Mode of transport**. State the mode of transport which you usually use for one journey to work or school." The wording of available check boxes in the question 22 was as follows:

- bus (other than city transport)
- mass city transport
- car – driver

- car – co-traveller
- train
- bicycle
- motorcycle
- other
- none (on foot only)

The **question 23** on the travel time had a following wording: "**Time spent on daily journey**. State how long one journey to work or school takes you." The wording of available check boxes in the question 23 was as follows:

- up to 14 min.
- 15 – 29 min.
- 30 – 44 min.
- 45 – 59 min.
- 60 – 89 min.
- 90 and more min.

The **question 24** on the journey frequency had a following wording: "**Frequency of journey to work or school**. State how often you commute from your municipality of residence to the municipality of your workplace." The wording of available check boxes in the question 24 was as follows:

- daily
- weekly
- 1 – 2 times per month
- other

The Census Form – Persons was accompanied by a document with self-explanatory name: **Explanations for the Census Form – Persons** (CSO , 2011b), which can be found in annex [A]. To be able to provide an exact definition of the modal split data, the wording of explanation related to question 21 through 24 will be presented in full:

"**Questions 21,22,23 and 24 on commuting to work or school** are filled out by persons who stated in the question on economic activity that they are employed or are pupils, students and apprentices (within the scope of economically inactive persons).
'Working students and apprentices' concerns information about commuting to school, not information about commuting to work. Commuting is also considered to be commuting within a town or city.

21. LOCATION OF PLACE OF WORK OR SCHOOL

State the address of the place of work (not, for example, the company headquarters).

Persons who have no workplace but start work at the same address (e.g. transport workers - drivers, pilots; artisans-repairers etc.), state where they start work.

Persons who often change their place of work (e.g. assembly and construction workers, marketplace sales people etc.), or travel but do not come to the same address daily (travelling businesspeople, taxi drivers, truck drivers etc.) should indicate 'employed without permanent workplace'.

Persons working or studying abroad should state the name of the state.

Persons whose workplace address (schools) is the same as the address of their (actual) residence, and employees without a permanent workplace do not fill out any further questions on commuting.

Questions 22 and 23 relate to daily journey to work or school (from home or from a place of temporary accommodation). Persons working in shifts enter information regarding a single journey to work even if they do not commute regularly every day.

22. MODE OF TRANSPORT

Bus (other than city transport) is stated by persons using bus transport which crosses the border of a municipality/city. It includes suburban transport.

Mass city transport is mass public transport operated for the satisfaction of the transport requirements for a city area. In Prague, it includes the metro.

Other means of transport includes all others, namely those not listed on the form.

None (on foot only) is stated by persons who walk to work or school, i.e. they do not use any means of transport.

23. TIME SPENT ON DAILY JOURNEY

State the duration of one normal journey to work or school. The time in minutes includes the total time which elapses from leaving home (or a place of temporary accommodation) to entry to the workplace or entry to school ('door to door'), i.e. including the walk to a public transport station and from a station, waiting for arrival, change etc.

24. FREQUENCY OF JOURNEY TO WORK OR SCHOOL

States how often a person commutes from their municipality of residence to the municipality of their workplace or school. Persons whose workplace or school is in the same town of (actual) residence do not answer the question.

Daily is stated by persons who commute directly from their place of (actual) residence and do not use temporary accommodation at the site of the workplace or school (rental, boarding house, student

hostel, school hostel, etc.). They may also be persons working in shifts or for a shortened time if they meet the stated conditions, even if they do not commute regularly every day."

The **answers** were collected and processed by the Czech Statistical Office (hereinafter CSO). The CSO offers more variants of the resulting files, which differ by the information value and the number of journeys included and described in the data files. The difference originates in the fact that not all questionnaires were completely answered by the respondents. Thus, the more information about the journeys is required, the lesser is the set of journeys, for which all the required information was provided. For instance, there is nothing else known about all the 3 920 953 journeys to work and 1 354 979 journeys to school, besides the simple fact these journeys were taken for these purposes. Those are the total numbers of journeys, which can be seen in the row 1 in Table 2. On the other hand, the subset on the row 2 consist of 2 024 876 work journeys and 775 901 school journeys, which are all supplemented by the information about their destination, at least. Besides the row 1, which includes all journeys without any further distinction, all other rows are including only the journeys, which had their origin and destination in Czechia (no travel abroad included). At the time of the national population census, there were 10 436 560 inhabitants residing in Czechia.

Table 2: Comparison of files available from CSO and produced within the research, Info = information value, W+O muni = within and outside the municipality, Out muni = outside the municipality, All freq's = all frequencies, Upon req = Upon request, Incompl R = incomplete records in terms of travel time and transport mode, Compl R = complete records

Commute data options from CSO					Data pre-processing methods			Journeys	
#	Info	Destination	Frequency	Availability	Selection	Reduction	Filtration	Work	School
1	Basic	Unknown	Unknown	Public	-	-	-	3 920 953	1 354 979
2	Extra	W+O muni	All freq's	Public	-	-	-	2 024 876	775 901
3	Extra	Out muni	All freq's	Public	-	-	-	1 099 928	421 683
4	Extra	Out muni	Daily	Upon req	Incompl R	National	Nonfilter	943 334	301 081
5	Extra	Out muni	Daily	Self-made	Compl R	National	Nonfilter	926 471	295 650
6	Extra	Out muni	Daily	Self-made	Compl R	Local	Nonfilter	528 413	164 353
7	Extra	Out muni	Daily	Self-made	Compl R	Local	Filtered	418 687	140 978

The data corresponding to the rows 1 through 3 of the Table 2 are publicly available. However, for the purpose of this research, it is critically important to distinguish the frequency of the travel. Such data set was obtained from the CSO upon a special request. The data set was provided in MS Excel format. The structure of the obtained data can be seen on the Figure 8 below. The CSO refers to this type of data set as a 'daily inter-municipal in- and out-commuting flows' or simply **'the flows'**. The number of journeys within Czechia included in the flows can be seen in the row 4 in Table 2 (emphasised by a red rectangle). However, the flows were provided including the redundant travel abroad. So, the original size of the flows (as obtained from CSO) was 447 640 data rows (records). The number of commuters

to work was 962 057, while the number commuters to school was 304 873. That in total makes together 1 266 930 commuter journeys.

	A	B	C	D	E	F	G	H	I
1	OBECOP_KOD	OKRESOP_KOD	OBECPS_KOD	OKRESPS_KOD	PROSTR	DOBA	ZAM	SKOLA	CELKEM
2	500011	40851	500496	40789	2	5	0	1	1
3	500011	40851	541630	40843	1	3	0	1	1
4	500011	40851	541630	40843	1	5	0	2	2
5	500011	40851	544787	40843	4	3	1	0	1
6	500011	40851	544990	40843	4	3	1	0	1
7	500011	40851	545058	40843	4	4	1	0	1
8	500011	40851	549401	40851	4	3	1	0	1
9	500011	40851	549622	40851	30	1	2	0	2
10	500011	40851	549622	40851	39	2	2	0	2
11	500011	40851	549622	40851	4	1	6	0	6
12	500011	40851	549622	40851	5	1	1	0	1
13	500011	40851	549622	40851	7	1	2	0	2
14	500011	40851	549622	40851	7	2	1	0	1
15	500011	40851	549622	40851	9	2	2	0	2

Figure 8: Preview of the original 'daily inter-municipal in- and out-commuting flows' data set from CSO

The meaning of the table heading is as follows:

OBECOP_KOD – code (ID) of the municipality of origin (out-commuting)

OKRESOP_KOD – code (ID) of the district of origin (out-commuting)

OBECPS_KOD – code (ID) of the municipality of destination (in-commuting)

OKRESPS_KOD – code (ID) of the district of destination (in-commuting)

PROSTR – mode of transport

DOBA – travel time

ZAM – number of commuters to work

SKOLA – number of commuters to school

CELKEM – number of commuters to work or school (total)

The **code of municipality** (and district) are unique and standardised ID used by the CSO for unambiguous identification of these administrative units across all the information resources of CSO.

The **mode of transport** is numbered according to specific code list. The list includes 94 values representing one of the 94 combinations of transport modes recognised by the CSO. The code list is discussed further in chapter 4.1.3.

The **travel time** is numbered according to specific code list. The list includes 7 values representing one of the 7 travel time intervals recognised by the CSO. The code list is discussed further in chapter 4.1.4.

The **number of commuters** is the same as the number of journeys, it is the number of individual persons making a single journey to work or school daily, leaving their municipality of residence (home or a place of temporary accommodation).

The key attribute of this data set (daily inter-municipal in- and out-commuting flows) is that **it is NOT YET structured** in a way one origin-destination pair (further abbreviated as O-D pair) would be described in a single data record (a row of the data set). On the contrary, the number of commuter journeys from one municipality to another (within one O-D pair) can be recorded in more rows if the transport mode or travel time differs. A new row is provided for every unique combination of transport mode and travel time used by the commuters in the O-D pair. For example, the rows 3 and 4 of the data (see Figure 8) are both related to commuting from the municipality 500011 to municipality 541630. The row 3 is showing the number of commuters, who stated their travel time is within the interval number 3, while the row 4 is showing the number of commuters, who stated their travel time is within the interval number 5. Similarly, the rows 9 through 15 are showing use of different transport modes (coded as 30, 39, 4, 5, 7 and 9) taken from the municipality 500011 to municipality 549622.

4.1.2 Data selection

The original 'daily inter-municipal in- and out-commuting flows' data set is not fully suitable for further analysis, because it consists **unknown values**. Typically, the code 999999 represents an unknown municipality, while the code 888888 represent a destination abroad. In regard to the travel time, the code 9 represents an unknown travel time, meaning it was not possible to abstract the travel time information from the census form. Similarly, the code 999 represents an unknown mode of transport, which was not possible to abstract. All the rows with any of the above-mentioned unknown values **were deleted** from the data set.

After deleting the incomplete records, the data set has shrunk to 419 516 rows and 1 222 121 commuting journeys, losing only about 3.4 % of the commuting journeys.

4.1.3 Aggregation of transport modes

As mentioned before, the question 22 in the census form offers 9 basic options for the transport mode. However, the form does not set a limit for their mutual combination. Thus, the interviewed citizens could select multiple transport modes to create **combinations of transport modes**. In the end, the CSO recognises 93 transport modes codes within the data set. (The transport mode code 999 was deleted it in the previous step, thus reducing the number of transport mode codes from 94 to 93.) The codes represent either one of the 9 single basic modes; or one of the 28 combination of 2 modes; or one of the 56 combinations of 3 modes. As the basic single modes are building stones for the

combinations, it is necessary to introduce the abbreviations for them to limit the length of the combined names. The abbreviations are presented in Table 3.

Table 3: List of abbreviations for the 9 basic transport modes

Transport mode	**Abbreviation**
Bus (other than city transport)	Bus
Train	Train
Mass City Transport	MCT
Car – Driver	Driver
Car – Co-traveller	CarPass
Motorcycle	Moto
Bicycle	Bike
Other	Other
None (on foot only)	Walk

To rationalise the modal split research, **it is necessary to aggregate** these 93 transport mode options into much smaller number of representative transport modes. As can be seen in the figures below, the 9 single basic modes are altogether having the largest modal share, while the combinations have a minor share. On the other hand, some of the single modes (typically the Moto) have much smaller share on its own then the aggregated share of the combinations. The figures below include the number of commuter journeys per each slice and their share on the total number of journeys.

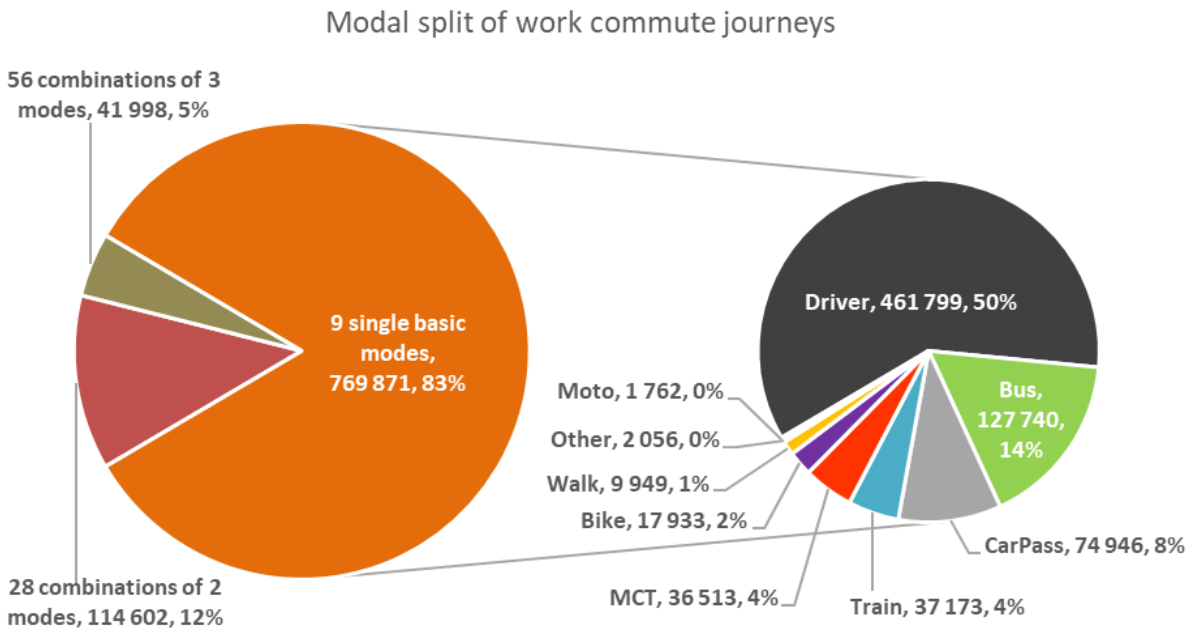


Figure 9: Modal split of work commute journeys before aggregation

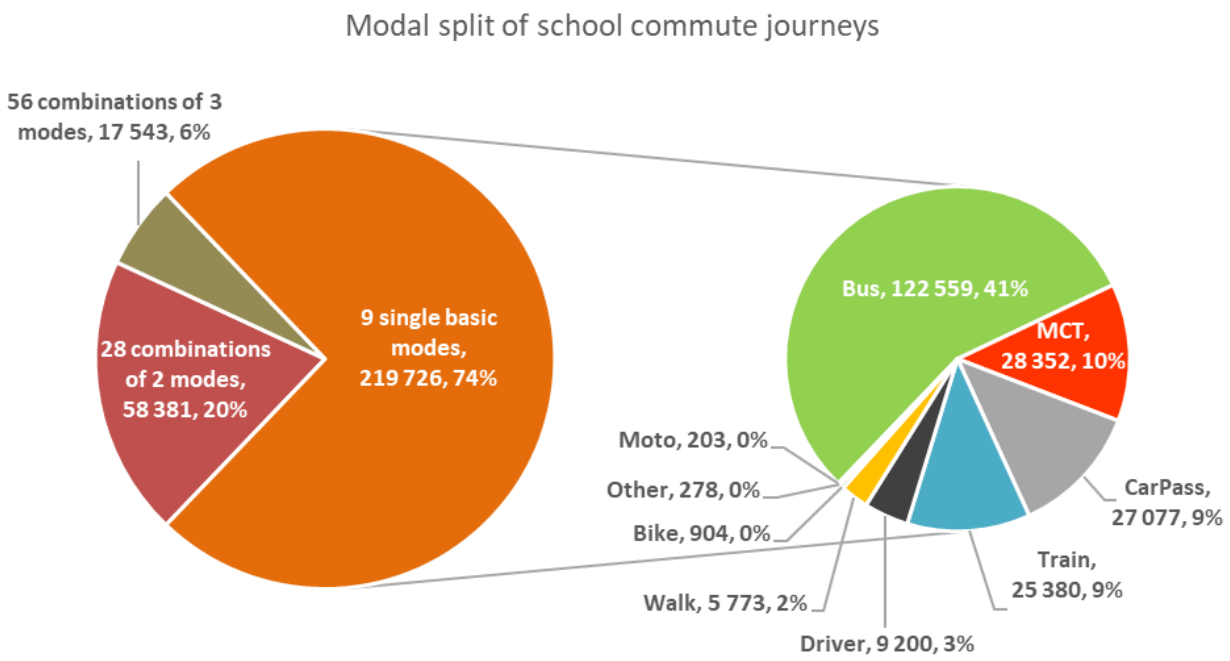


Figure 10: Modal split of school commute journeys before aggregation

The **aim of the aggregation** is to create a set of representative transport modes, which will have more even distribution of the modal share while distinguishing all specific transport options. All 93 options were included into the aggregation process, so no journeys were lost. This is the advantage over the methodologies, where multiple transport options with negligible share are omitted, despite those have a significant share altogether. The proposed aggregation scheme is presented in Table 4 below. The original 93 options on the left are aggregated to 10 new modes on the right. Commuter journeys to work and to school are presented separately.

Table 4: Scheme for aggregation of 93 original transport modes into 10 new modes

93 ORIGINAL TRANSPORT MODES				→	**10 NEW AGGREGATED TRANSPORT MODES**		
#	Abbreviation	Work	School		**#**	**Abbreviation**	**Work**	**School**
1	Bus	127 740	122 559		**1**	**Bus**	**127 860**	**122 662**
16	Bus+Other	120	103					
2	Train	37 173	25 380		**2**	**Train**	**37 252**	**25 424**
22	Train+Other	79	44					
3	MCT	36 513	28 352		**3**	**MCT**	**36 626**	**28 427**
27	MCT+Other	113	75					
4	Driver	461 799	9 200		**4**	**Driver**	**466 435**	**9 481**
6	Moto	1 762	203					
29	Driver+Moto	2 719	73					
31	Driver+Other	123	3					
36	Moto+Other	8	2					
88	Driver+Moto+Other	24	0					
5	CarPass	74 946	27 077		**5**	**CarPass**	**74 998**	**27 104**
34	CarPass+Other	52	27					

93 ORIGINAL TRANSPORT MODES				→	**10 NEW AGGREGATED TRANSPORT MODES**		
#	Abbreviation	Work	School		**#**	**Abbreviation**	**Work**	**School**
12	Bus+Driver	12 822	1 101		**6**	**PTplus**	**76 364**	**24 437**
13	Bus+CarPass	9 669	7 629					
14	Bus+Moto	300	270					
15	Bus+Bike	3 709	1 002					
18	Train+Driver	4 399	407					
19	Train+CarPass	1 664	882					
20	Train+Moto	76	57					
21	Train+Bike	1 675	330					
23	MCT+Driver	5 407	639					
24	MCT+CarPass	2 778	2 029					
25	MCT+Moto	55	43					
26	MCT+Bike	973	205					
39	Bus+Train+Driver	2 782	536					
40	Bus+Train+CarPass	1 717	1 573					
41	Bus+Train+Moto	63	96					
42	Bus+Train+Bike	840	388					
44	Bus+MCT+Driver	4 349	1 087					
45	Bus+MCT+CarPass	2 469	2 066					
46	Bus+MCT+Moto	45	73					
47	Bus+MCT+Bike	405	143					
49	Bus+Driver+CarPass	5 145	589					
50	Bus+Driver+Moto	528	80					
51	Bus+Driver+Bike	3 350	81					
52	Bus+Driver+Other	46	4					
53	Bus+CarPass+Moto	141	154					
54	Bus+CarPass+Bike	1 356	367					
55	Bus+CarPass+Other	52	31					
56	Bus+Moto+Bike	112	29					
57	Bus+Moto+Other	1	1					
58	Bus+Bike+Other	44	19					
59	Train+MCT+Driver	2 839	703					
60	Train+MCT+CarPass	1 082	942					
61	Train+MCT+Moto	27	37					
62	Train+MCT+Bike	679	285					
64	Train+Driver+CarPass	964	109					
65	Train+Driver+Moto	170	19					
66	Train+Driver+Bike	959	33					
67	Train+Driver+Other	33	3					
68	Train+CarPass+Moto	33	17					
69	Train+CarPass+Bike	216	44					
70	Train+CarPass+Other	13	11					
71	Train+Moto+Bike	31	4					
72	Train+Moto+Other	3	3					
73	Train+Bike+Other	20	7					
74	MCT+Driver+CarPass	947	139					

93 ORIGINAL TRANSPORT MODES				→	**10 NEW AGGREGATED TRANSPORT MODES**		
#	Abbreviation	Work	School		**#**	**Abbreviation**	**Work**	**School**
75	MCT+Driver+Moto	157	27				
76	MCT+Driver+Bike	847	46				
77	MCT+Driver+Other	29	3				
78	MCT+CarPass+Moto	22	27				
79	MCT+CarPass+Bike	246	46				
80	MCT+CarPass+Other	20	7				
81	MCT+Moto+Bike	27	3				
82	MCT+Moto+Other	1	0				
83	MCT+Bike+Other	27	11				
7	Bike	17 933	904		**7**	**Bike**	**18 002**	**910**
37	Bike+Other	69	6				
10	Bus+Train	8 037	11 369		**8**	**PTcom**	**52 786**	**50 604**
11	Bus+MCT	22 754	18 259				
17	Train+MCT	15 927	13 326				
38	Bus+Train+MCT	5 830	7 431				
43	Bus+Train+Other	52	39				
48	Bus+MCT+Other	94	75				
63	Train+MCT+Other	92	105				
9	Walk	9 949	5 773		**9**	**Walk**	**9 949**	**5 773**
8	Other	2 056	278		**10**	**Rest**	**26 199**	**828**
28	Driver+CarPass	11 761	306				
30	Driver+Bike	7 761	40				
32	CarPass+Moto	187	28				
33	CarPass+Bike	1 241	118				
35	Moto+Bike	124	8				
84	Driver+CarPass+Moto	455	22				
85	Driver+CarPass+Bike	1 427	9				
86	Driver+CarPass+Other	50	2				
87	Driver+Moto+Bike	943	7				
89	Driver+Bike+Other	108	1				
90	CarPass+Moto+Bike	61	2				
91	CarPass+Moto+Other	1	1				
92	CarPass+Bike+Other	16	6				
93	Moto+Bike+Other	8	0				

The **key ideas behind the** proposed **aggregation** scheme are as follows:

1) The combinations of any transport mode with only transport option Other was aggregated with the particular transport mode as it is the decisive one. This way, the traditional modes Bus, Train, MCT, Driver, CarPass and Bike were enlarged by their combination with mode Other. However, this enlargement should have no significant impact on the structure of these modes as the combinations with mode Other are in every case a fraction of percent.

2) The sparsely used transport mode Moto was aggregated with mode Driver. The mode Moto simply has not the modal share large enough compared to the other modes. Modes Driver and Moto are both private, motorised modes of transport, which gives the ground for their aggregation. It should not have any significant impact on the structure of the mode Driver as the mode Moto makes only about 1 % of it in the case of work commute and about 3 % of it in the case of school commute.
3) There was no transport mode suitable for aggregation with the transport mode Walk, so it was not aggregated in any way.
4) Two new transport mode consisted of many various combinations were created.
    a. **PTplus** – combinations of public transport modes (Bus, Train, MCT) with private transport modes (Driver, Moto, CarPass, Bike). This mode represents the option, where the private transport is used for easing the access to the public transport from the origin, or to the final destination from the public transport stop. Thus, the public transport (**PT**) has a benefit (**plus**) from the private transport.
    b. **PTcom** – **com**bination of only public transport (**PT**) modes (Bus, Train, MCT).
5) The remaining combinations and the single mode Other were aggregated into new mode named 'Rest' to avoid confusion with the original mode Other. However, the purpose of the mode Rest is the same as in the case of the mode Other – to aggregate the sparsely used modes, which do not have the modal share large enough for independent study and which do not show similarities with any other mode, but need to be kept in the computation base of the modal split to avoid its distortion.

The final modal split after the aggregation of the 93 original transport mode options into 10 new aggregated modes can be seen in the figures below.

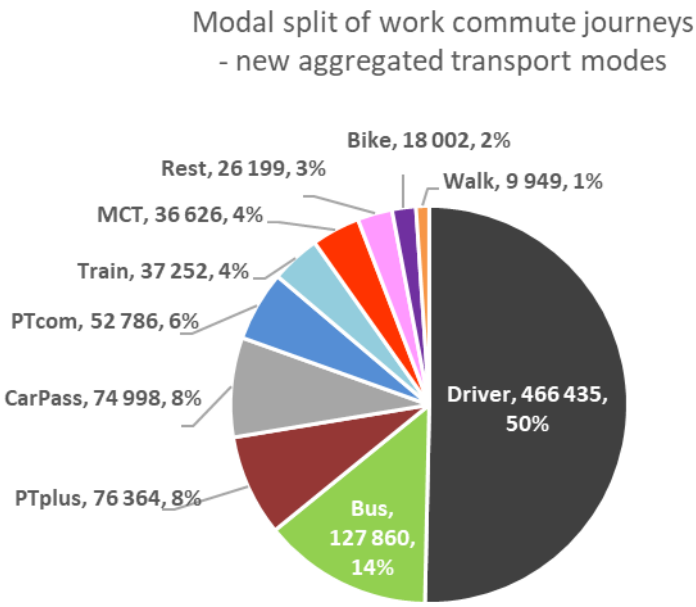


Figure 11: Modal split of work commute journeys after aggregation

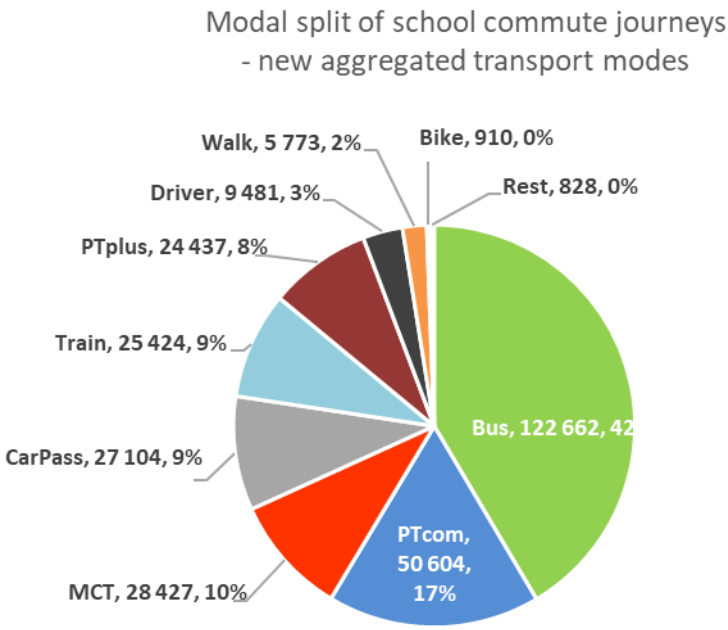


Figure 12: Modal split of school commute journeys after aggregation

The numbers of aggregated modes 1 through 10 were then added to the original 'flows' data set and will play a role during the transformation of the O-D pairs described in chapter 4.1.5.

4.1.4 Preparation for travel time averaging

As mentioned in chapter 4.1.1, the travel time in the original data set is recorded in a form of six codes, each representing a given time interval of travel times from the census form. The structure of the original 'flows' data set can include multiple different travel time codes in one O-D pair per one transport mode. To obtain a single travel time per O-D pair and one transport mode, the travel times

need to be averaged. But before it can be done so, it is necessary **to replace the travel time intervals** (represented by the codes) with interval mean time, reflecting the varying length of the intervals. The overview of travel intervals, travel time codes and proposed interval means can be seen in Table 5.

Table 5: Overview of travel time averaging parameters

Travel time interval code	Travel time interval	Interval mean time	Boundary between intervals	Average of the adjacent interval means	Difference
1	up to 14 min.	7.5 min.			
			15 min.	15 min.	0 min.
2	15 – 29 min.	22.5 min.			
			30 min.	30 min.	0 min.
3	30 – 44 min.	37.5 min.			
			45 min.	45 min.	0 min.
4	45 – 59 min.	52.5 min.			
			60 min.	63.5 min.	3.5 min.
5	60 – 89 min.	74.5 min.			
			90 min.	97 min.	7 min.
6	90 and more min.	119.5 min.			

The **key ideas behind the interval mean time selection** are as follows:

1) Chose the mean as close as possible to the centre of the interval
2) Chose the mean in a way the average of the two adjacent interval means would be equal to the boundary between the intervals in the case of the intervals of the same length
3) Choose the mean of the longer intervals in a way the resulting average mean of the adjacent intervals will be closer to the mean of the longer interval
4) Choose the mean in a way the difference between the boundary of the adjacent intervals and the average of their means gradually increases with prolonging of the longer intervals

Once the travel time codes in the original 'flows' data set are replaced by the interval mean times, it is than possible **to average the travel times**, which are related to the same O-D pair and the same transport mode. The interval mean time added to the original 'flows' data set will play a role during the transformation of the O-D pairs described in chapter 4.1.5.

4.1.5 Transformation of origin-destination pairs

The Aggregation of the transport modes and the travel time averaging were both a **prerequisite** for the transformation as they added important imputes to the original 'flows' data set. As can be seen in Figure 13, the column D was now added to the data set as a result of transport mode aggregation. Moreover, the column E, which is a result of the travel time averaging, has replaced the column with the travel time codes, which are no longer needed.

	A	B	C	D	E	F	G	H
1	ORIGIN_MUNI	DESTIN_MUNI	MODE 93	AGR 10	MIN	WORK	SCHOOL	TOTAL
2	500011	500496	2	2	74.5	0	1	1
3	500011	541630	1	1	37.5	0	1	1
4	500011	541630	1	1	74.5	0	2	2
5	500011	544787	4	4	37.5	1	0	1
6	500011	544990	4	4	37.5	1	0	1
7	500011	545058	4	4	52.5	1	0	1
8	500011	549401	4	4	37.5	1	0	1
9	500011	549622	30	10	7.5	2	0	2
10	500011	549622	39	6	22.5	2	0	2
11	500011	549622	4	4	7.5	6	0	6
12	500011	549622	5	5	7.5	1	0	1
13	500011	549622	7	7	7.5	2	0	2
14	500011	549622	7	7	22.5	1	0	1
15	500011	549622	9	9	22.5	2	0	2

Figure 13: Preview of the 'flows' data set with new columns D and E

The meaning of the table heading is as follows:

ORIGIN_MUNI – code (ID) of the municipality of origin (out-commuting)

DESTIN_MUNI – code (ID) of the municipality of destination (in-commuting)

MODE 93 – 93 original transport modes

AGR 10 – 10 new aggregated transport modes

MIN – averaged travel time in minutes

WORK – number of commuters to work

SCHOOL – number of commuters to school

TOTAL – number of commuters to work or school

The goal of the transformation is to organise all the records related to **one O-D pair to one row** of the data set. Example can be seen in the comparison of Figure 13 and Figure 14. The transport mode and travel time information regarding the O-D pair from the municipality 500011 to the municipality 549622 was in the Figure 13 recorded in seven rows number 9 through 15 (marked by a red rectangle), while the same information is in the Figure 14 solely in row number 9. The Figure 14 only shows half of the transformed row with the work commute related information, because the O-D pair from the municipality 500011 to the municipality 549622 does not include any journey to school. The structure of the transformed data set allows simple split of the data set into work commute and school commute part later on.

	L	M	N	O	P	Q	R	S	T	U	V	W	BE	BF	BG	BH	BI	BJ	BK	BL	BM	BN
1	Municipalities		Number of journeys to work by transport mode (AGR 10)										Average travel time to work by transport mode (MIN)									
2	Origin	Destin	Bus	Train	MCT	Driver	CarPass	PTplus	Bike	PTcom	Walk	Rest	Bus	Train	MCT	Driver	CarPass	PTplus	Bike	PTcom	Walk	Rest
9	500011	549622	0	0	0	6	1	2	3	0	2	2				7.5	7.5	22.5	15		22.5	7.5

Figure 14: Preview of the transport mode and travel time data regarding the O-D pair from the municipality 500011 to the municipality 549622 organised into one row

The transformation of the transport mode information is based on **summing the number of journeys** taken by each of the 10 aggregated modes over all records related to the respective O-D pair. The transformation of the travel time information is based on **averaging the travel times** by each of the 10 aggregated modes over all records related to the respective O-D pair.

As a **result of the transformation**, information about each O-D pair is recorded only in one data row and vice versa. Thus, information about every O-D pair is recorded in a data space of the same size, which significantly eases the following analyses. After the transformation, the number of the O-D pairs and the number of journeys in the data set remain unchanged. The difference is in the number of recognised transport modes, which has decreased from 93 to 10 (transport mode aggregation); in averaging the travel time per each of the 10 aggregated modes in one O-D pair; and in the number of data rows in the data set, which has decreased from 419 516 to 126 902. The latter one is also the number of O-D pairs in the data set.

4.1.6 Reduction of origin-destination pairs

The **need for a reduction** of the data set by some O-D pairs is given by the primary focus of the research, which was specifically set to the local commute. Based on the literature review, the municipalities with extended powers (hereinafter MEP) are the natural centres of the local commute in the Czech environment. The commuting catchment areas of the superior regional capitals as well as the national capital are much broader, having partially character of long-distance travel. Thus, O-D pairs to and from regional capitals and the Capital City of Prague were removed from the data set. To further diminish the journeys on longer distances, the mutual connections of the MEPs were also removed. This should keep in the data set predominantly the journeys to the local commute centres from their commuting hinterland and vice versa.

A preparation for the reduction involves adding the code for a level of government to every municipality. The MEPs are in the Czech governmental hierarchy labelled as the level 3. The regional capitals are level 4; and the Capital City of Prague is the only 5. Reduction was then practically done by **deleting the O-D pairs** (data rows), which are from, to or between the municipalities (cities) of level 4 or 5; and then deleting the O-D pairs between the municipalities (towns) of level 3.

As a **result of the reduction**, the number of the origin-destining pairs has decreased from 126 902 to 105 577, the number of work commute journeys has decreased from 926 471 to 528 413 and similarly the number of school commute journeys from 295 560 to 164 353. The vast number of deleted journeys is a pity, however it is given by the high average number of travels within O-D pairs to and from the national and regional capitals, which were excluded from the local level, if you wish, from the rural level commute.

4.1.7 Filtration of origin-destination pairs

The **introduction to filtration** was made in a previous work of Šatra and Čarský (2020) on the filtration of O-D pairs in the following matter: "After the initial reduction, the starting data set for our research is including O-D pairs from 193 MEPs to municipalities on subordinate levels and vice versa plus O-D pairs between the municipalities on the subordinate levels. This data set will be further referred as "Nonfiltered". It includes all the O-D pairs on the required local level. (…) All the 193 included MEPs are having status of town, ranging approx. from 2 800 to 76 700 inhabitants with average population size of approx. 16 000 inhabitants. (…) Average area size of administrative district of MEP in the Czech Republic is about 385 km^2."

The **'Nonfiltered' data set**, which was the result of the reduction is consisted of 105 577 O-D pairs. However, 55 505 of them are pairs with only one traveller, which is over 52 % of all O-D pairs. In other words, in the case of 55 505 municipality pairs, there is only one traveller between them in one direction. "Whichever transport mode the one traveller choses, will be the only transport mode used within that O-D pair. Then the modal split, describing the distribution of shares of transport modes within this O-D pair, is only consisted of the one used transport mode, which has share of 100 % and all the other modes have share of 0 %. This is an extreme distortion of the modal split in favour of only one transport mode.

Considering the number of O-D pairs with one traveller causing the distortion, one can consider the whole data set to be misleading about the overall modal split. The solution is **to remove the O-D pairs of lesser importance** from the data set. It is assumed that most of the O-D pairs with one traveller are of lesser importance. Similarly, all other O-D pairs with small number of travellers, but the O-D pairs with one traveller were chosen to be the monitored key performance indicators as they cause the greatest distortion.

But **how to assess the importance of the O-D pairs**? Which O-D pairs must stay in the data set to keep its informative value? Starting with the latter one, the requirements were set as follows:

A. As this research should serve to the municipal and regional level of government, first of the requirements is to sustain all the municipalities in the data set. In other words, there should not be a municipality, whose O-D pairs will be all removed from the data set and so the municipality with them.
B. Second requirement is to maximize the number of the travellers in the data set. This is driven by the will to conduct a research, which will have an impact to maximum of travellers. Practically, this is about minimizing the loss of travellers from the data set caused by removing of O-D pairs.

Simple removing all O-D pairs with one traveller does not meet the criterion of sustaining all the municipalities in the data set because it results in removing 13 municipalities from it. This is due to the **specific municipal structure of the Czech Republic**.

At the time of the census, there were 6251 municipalities in the country, ranging from 17 to 1 268 796 inhabitants, having on average 1 670 inhabitants. Since the **municipalities can have** so **small population** as 17 inhabitants, it is then expectable, the O-D pairs originating or terminating in such municipality will have only one traveller. At the same time, the O-D pair with one traveller is an important O-D pair for this municipality as it was used by over 5 % of population of municipality and perhaps by more than 10 % of the workers and pupils residing in it. This gives the preview how the importance of the O-D pairs can be assessed." (ŠATRA and ČARSKÝ, 2020)

Before the filtration, the 'Nonfiltered' data set was divided into **two separate data sets**, first containing only O-D pairs of work commute and the other one with only O-D pairs of school commute. Some municipal pairs are in both data set, but there are also pairs of municipalities, between which only work or only school journeys are taken. For example, the above-mentioned O-D pair from the municipality 500011 to municipality 549622 does not include any journey to school, so it is included only in the journey to work data set.

The new **'Nonfiltered work'** data set consists of 98 084 O-D pairs from 6 232 municipalities of origin to 5 587 municipalities of destination and it includes 528 413 journeys in total. The new **'Nonfiltered school'** data set is smaller. It consists of only 28 100 O-D pairs from 6 171 municipalities of origin to 2 344 municipalities of destination and it includes 164 353 journeys in total.

4.1.7.1 T_{ij} filtration method

The theoretical background for the **first filtration method** was set by Afonso and Venâncio (2016), who propose the following formula for assessment of the O-D pair importance:

$$T_{ij} = \frac{C_{ij}}{\min(r_i, r_j)} \tag{18}$$

where:

T_{ij} – strength of commuting tie between two municipalities i and j

C_{ij} – number of travellers (workers or school attendees) residing in the municipality i and commuting to municipality j to work or school

r_i – number of all workers or school attendees residing in municipality i

r_j – number of all workers or school attendees residing in municipality j

"The **parameter T_{ij}** can be used for assessment of the O-D pair importance. Afonso and Venâncio recommend considering the O-D pairs with T_{ij} above 0.02 to be 'stronger' (important from the perspective of this research).

For **example**, if there would be a municipality i, where in total 100 workers are residing and two travellers are commuting from there to municipality j, where in total 500 workers are residing. Then, the C_{ij} is 2 and smaller of the two r is r_i, which is 100. By dividing 2/100 we get that the T_{ij} of this O-D pair is 0.02; just below the set criteria, thus it is an unimportant O-D pair sentenced to be removed from the data set." (ŠATRA and ČARSKÝ, 2020)

However, the **experience with using the T_{ij} filtration method** is that the level of importance of 0.02 (i.e. 2 %) is overly strict, removing too much of O-D pairs and even some municipalities with them. This means, less strict threshold needs to be found to sustain all the municipalities in the data set. This is practically done by finding the maximum T_{ij} for each municipality over all its O-D pairs, which start or terminate in the municipality, and then the threshold will be the least of all the municipal T_{ij} maximums. This approach can be applied separately to origin municipalities and destination municipalities. The destination municipalities tend to have smaller T_{ij} importance level, thus, to keep all origin municipalities and destination municipalities in the data set, the threshold from destination municipalities should be used. For the work commute data set, the resulting T_{ij} threshold is 0.001709402 (0.17 %). In terms of the school data set, the threshold is 0.000493097 (0.05 %). The performance of the T_{ij} filtration method can be seen in Table 6 and Table 7. The tables also show comparison with the V_{ij} filtration method introduced further.

Table 6: Key attributes of the data sets (Files) related to filtration of O-D pairs of work commute, O-D pair = origin-destination pair, %↓ = decrease in %, number of Travellers = number of Journeys, Pw1T = O-D pair with only 1 Traveller, Pw1T/T = ratio of number of O-D pairs with only 1 Traveller to the total number of Travellers, Pw1T/P = ratio of number of O-D pairs with only 1 Traveller to the total number of O-D pairs, Nationwide work = data set of work commute journeys before reduction of O-D pairs to local commute

File \ Parameter	O-D Pairs	%↓	Travellers	%↓	Pw1T	%↓	Pw1T/T	Pw1T/P
Nationwide work	116 069		926 471		60 702		6.55%	52.30%
Nonfiltered work	98 084	base	528 413	base	53 484	base	10.12%	54.53%
T_{ij} filtered work	82 200	-16.2%	509 271	-3.6%	39 798	-25.6%	7.81%	48.42%
V_{ij} filtered work	37 406	-61.9%	418 687	-20.8%	11 333	-78.8%	2.71%	30.30%

Table 7: Key attributes of the data sets related to filtration of O-D pairs of school commute, see Table 6 for legend

File \ Parameter	O-D Pairs	%↓	Travellers	%↓	Pw1T	%↓	Pw1T/T	Pw1T/P
Nationwide school	40 253		295 650		16 953		5.73%	42.12%
Nonfiltered school	28 100	base	164 353	base	11 700	base	7.12%	41.64%
T_{ij} filtered school	27 451	-2.3%	163 673	-0.4%	11 076	-5.3%	6.77%	40.35%
V_{ij} filtered school	14 731	-47.6%	140 978	-14.2%	2 654	-77.3%	1.88%	18.02%

The T_{ij} filtration method did filter out much larger number of O-D pairs with only 1 traveller from the work commute data set than in the case of the school commute data set (25.6 % over 5.3 %), however, the large number of these remain in both data sets, about 48 % of all O-D pairs in the work commute data set and about 40 % in the school commute data set. If **nearly half of the data set is still distorted** (because of 100 % share of a transport mode in O-D pairs with only one traveller), there is a ground to consider the whole data set misleading. This stimulates the need to develop another filtering method, which would remove larger number of O-D pairs with only one traveller.

4.1.7.2 V_{ij} filtration method

"The T_{ij} method focuses on the assessing the importance of the O-D pairs based on the proportion of the travellers using the O-D pairs to the travellers (workers or school attendees) residing in the smaller of two municipalities. The drawback is that it does not deal with the proportion of the travellers using the O-D pair to the total number of travellers leaving the municipality of origin to work or school." (ŠATRA and ČARSKÝ, 2020) This can be changed by **introducing the parameter V_{ij},** which takes into consideration the total number of travellers leaving the municipality of origin for the particular purpose (work or school) as well as the total number of travellers entering the municipality of destination for one of the purposes. The parameter V_{ij} was designed as an analogy to T_{ij} with a following formula:

$$V_{ij} = \frac{C_{ij}}{\min(\sum_1^m C_{ix}, \sum_1^n C_{yj})} \tag{19}$$

where:

V_{ij} – strength of commuting tie between two municipalities i and j

C_{ij} – number of travellers (workers or school attendees) residing in the municipality i and commuting to municipality j to work or school

C_{ix} – number of travellers (workers or school attendees) residing in the municipality i and commuting to municipality x to work or school, where $x = (1 \dots m)$

1 – the first O-D pair from municipality i

m – the last O-D pair from municipality i

C_{yj} – number of travellers (workers or school attendees) residing in the municipality y and commuting to municipality j to work or school, where $y = (1 \dots n)$

1 – the first O-D pair to municipality j

n – the last O-D pair to municipality j

The **threshold for the V_{ij} importance level**, which would keep all the origin municipalities and destination municipalities in the data set, can be found in the same way as in the case of the T_{ij} parameter. For the work commute data set, the resulting V_{ij} threshold is 0.0667 (6.67 %). In terms of the school data set, the threshold is 0.125 (12.5 %).

The **performance of the V_{ij} filtration method** can be seen in Table 6 and Table 7. Based on the comparison with the T_{ij} filtration method, the V_{ij} method has removed much larger number of O-D pairs with only one traveller from the data set. Moreover, the number of removed pairs is similar for both data set of work commute and school commute, about 79 %, 77 % respectively. The drawback of the V_{ij} filtration method is that removes more of the O-D pairs with multiple travellers and thus large portion of all the travellers from the data set. However, the percentage decrease in the total number of O-D pairs as well as the percentage decrease in the number of travellers in the data set are smaller than the percentage decrease in the number of O-D pairs with only one traveller. Another control attribute is the ratio of number of O-D pairs with only 1 traveller to the total number of travellers (Pw1T/T), which is in the case of V_{ij} filtration method significantly better than in the case of the T_{ij} method. This attribute in fact specifies, how many percent of all travellers are travelling on all O-D pairs with only 1 traveller. This attribute is also favouring the V_{ij} filtration method.

The **resulting number of journeys** remaining in the work and school commute data sets after the application of the V_{ij} filtration method can be seen in in the row 7 of the Table 8.

Table 8: Comparison of files available from CSO and produced within the research, Info = information value, W+O muni = within and outside the municipality, Out muni = outside the municipality, All freq's = all frequencies, Upon req = Upon request, Incompl R = incomplete records in terms of travel time and transport mode, Compl R = complete records

Commute data options from CSO					Data pre-processing methods			Journeys	
#	**Info**	**Destination**	**Frequency**	**Availability**	**Selection**	**Reduction**	**Filtration**	**Work**	**School**
1	Basic	Unknown	Unknown	Public	-	-	-	3 920 953	1 354 979
2	Extra	W+O muni	All freq's	Public	-	-	-	2 024 876	775 901
3	Extra	Out muni	All freq's	Public	-	-	-	1 099 928	421 683
4	Extra	Out muni	Daily	Upon req	Incompl R	National	Nonfilter	943 334	301 081
5	Extra	Out muni	Daily	Self-made	Compl R	National	Nonfilter	926 471	295 650
6	Extra	Out muni	Daily	Self-made	Compl R	Local	Nonfilter	528 413	164 353
7	Extra	Out muni	Daily	Self-made	Compl R	Local	Filtered	418 687	140 978

The total number of journeys remaining for further research within this thesis is 559 665 journeys to work or school. That is about 54 journeys per 1 000 inhabitants. On the other hand, the **Czechia in motion** project, which was conducting the mobility survey of selected households, is working with 51 434 individual trips, not journeys. Since the trips can be further chained into complex journeys, the number of journeys in Czechia in motion is most likely even smaller. In any case, if dividing the number of trips by the number of inhabitants, there was about 5 trips per 1 000 inhabitants.

The total number of O-D pairs remaining in the work commute data set is 37 406, thus the average number of journeys in one O-D pair is 11.19. In the school commute data set remained 14 731 O-D pairs and the average number of journeys in one O-D pair is 9.57.

4.1.8 Modal split calculation

The last step of data pre-processing which directly precedes any further modal split analysis or modelling is a calculation of modal split of O-D pairs. The Figure 15 is providing an example of a calculated modal split of the pair from the municipality 500011 to the municipality 549622, which have been used as an example in other pre-processing methods too. The Figure 15 also provides a preview of the calculation process. The modal share of each transport mode is calculated as a number of journeys taken by this mode (6 in the case of Driver) divided by the total number of journeys taken within the O-D pair (C_{ij} = 16). The resulting modal share of transport mode Driver (0.375) can also be read as a share of 37.5 %. The sum of shares of all transport mode gives 1 (100 %); and the set of all mode shares is the modal split of the O-D pair.

W3 | =L3/$S3

	C	D	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC
1	**Municipalities**		**Number of journeys to work by transport mode**										C_{ij}	**Modal split of the origin-destination pair**									
2	Origin	Destin	Bus	Train	MCT	Driver	CarPass	PTplus	Bike	PTcom	Walk	Rest		Bus	Train	MCT	Driver	CarPass	PTplus	Bike	PTcom	Walk	Rest
3	500011	549622	0	0	0	6	1	2	3	0	2	2	16	0	0	0	0.375	0.0625	0.125	0.19	0	0.13	0.13

Figure 15: Preview of modal split data of the O-D pair from the municipality 500011 to the municipality 549622

Note that the O-D pair from the municipality 500011 to the municipality 549622 is now in row 3 but used to be in a row 9 in the previous example. This is because the pairs, which used to be in rows 3-8 in the previous example are now included in the school commute data set only or were removed entirely during the filtration process (V_{ij} filtration method).

The exact **definition of the above presented modal split** is as follows: modal split of O-D pair from the municipality 500011 to the municipality 549622 calculated based on all daily journeys to work recorded within the O-D pair. It could also be added that this O-D pair is a part of the local commute and has been found important by the V_{ij} filtration method (importance level V_{ij} > 0.0667).

4.2 Collection and pre-processing of explanatory variables data set

The modal split dependency research will be based on models described in detail in chapter 4.3. About the models in short. The modal split will be the explained variable (dependent or Y variable), while the various **characteristics of either municipalities** or **the connections between them** will be the explanatory variables (independent or X variables).

As already indicated, **municipality-related models** will work with explanatory variables related to the municipalities of origin or destination such as demography, car ownership, real estate, geomorphology or availability of transport services in the municipality. The **connection-related models** will process the explanatory variables related to the connections such as travel distance, travel time or frequency of the transport services within the connection.

The explained variable (modal split) was collected to the decisive moment of the population census, which was on the midnight from 25. 03. to 26. 03. 2011. To search the dependency of the explained variable on a **time relevant explanatory variable**, it is desirable to obtain the explanatory variables collected or valid to date as close as possible to the decisive moment. Most of the explanatory variables are also coming from the census, so their time match is perfect. The others are to some other date during 2011 (typically 01. 01. 2011) or represent the overall data for all year 2011. Exceptions are described and explained further.

4.2.1 Variables related to municipalities

The list of all municipality-related explanatory variables used in the municipality-related models of modal split dependency can be seen in Table 9. The final list contains **39 explanatory variables**. The table distinguishes different demands on data pre-processing (preparation) and availability of the data sources. The variables marked as 'taken' were extracted from the source as they were and assigned to each municipality. The variables marked as 'calculated' required some processing (typically summing, averaging or multiplying) based on the data in the referred sources. Specific calculation details are described in chapter 4.2.2.2. The variables marked as 'generated' involved either a manual generation of new data set based on an internet research or a GIS assisted creation of new shape files and a subsequent export to table files. Specific details on the data generation are described in chapter 4.2.1.2. Regarding the data availability, some data sources were available only upon request, some even had to be officially purchased, but the majority of the data are available online.

The **green colouring** marks the variables, which were added to the set of explanatory variables in comparison to the Table 1 in chapter 3.7.1. Further comments on the added variables can be found in

chapter 4.2.1.3. The explanatory variables, which were omitted in comparison to the Table 1 are obviously not recorded in Table 9, but will be commented in chapter 4.2.1.4.

Table 9: List of all municipality-related explanatory variables used in the municipality-related models (A and B)

Code	Group	Explanatory variable	Preparation	Available	Source tag
x100	Age structure	Average age of inhabitants	taken	online	CSO_6
x110	Age structure	Share of children aged 0-14	taken	online	CSO_6
x120	Age structure	Share of inhabitants aged 15-64	taken	online	CSO_6
x130	Age structure	Share of inhabitants with age 65+	taken	online	CSO_6
x140	Age structure	Age index (x130/x110)	calculated	online	CSO_6
x200	Education	Share of employees with elementary education	taken	online	CSO_8
x210	Education	Share of employees with apprenticeship certificate	taken	online	CSO_8
x220	Education	Share of employees without high school diploma	calculated	online	CSO_8
x230	Education	Share of employees with high school diploma	taken	online	CSO_8
x240	Education	Share of employees with completed extension study	taken	online	CSO_8
x250	Education	Share of employees with tertiary technical education	taken	online	CSO_8
x260	Education	Share of employees with university degree	taken	online	CSO_8
x270	Education	Share of employees with high school diploma or higher	calculated	online	CSO_8
x280	Education	Average length of the education attendance of employees	calculated	online	CSO_8
x300	Econ. activity	Unemployment rate	taken	online	CSO_7
x310	Econ. activity	Share of economically active inhabitants	taken	online	CSO_7
x320	Econ. activity	Share of the retired	taken	online	CSO_7
x330	Econ. activity	Share of pupils, students and apprentices	taken	online	CSO_7
x340	Econ. activity	Ratio of economically inactive to active inhabitants	calculated	online	CSO_7
x350	Econ. activity	Ratio of self-employed to economically active inhabitants	calculated	online	CSO_9
x360	Econ. activity	Ratio of occupied jobs to economically active inhabitants	calculated	request	CSO_5
x370	Econ. activity	Ratio of inhabitants employed in key sectors (T+/PS-)	calculated	online	CSO_8
x400	Tax revenue	Share of municipal inhabitant on national tax revenue	calculated	online	MF_14
x410	Car ownership	Car ownership	calculated	online	MT_15
x420	Population size	Number of inhabitants	taken	online	CSO_7

Code	Group	Explanatory variable	Preparation	Available	Source tag
x430	Population size	Total number of workers, pupils, students and apprentices	calculated	online	CSO_7
x500	Pop. density	Gross population density	calculated	online	CSO_11
x510	Pop. density	Net population density	calculated	online	CSO_11
x600	Trans. services	Municipality organizing Mass City Transport	generated	online	Internet
x610	Trans. services	Distance to the closest train stop or station	generated	request	ArcData_1
x700	Real estate	Average purchase price of house	taken	request	CSO_2
x710	Real estate	Average purchase price of apartments	taken	request	CSO_3
x720	Real estate	Average purchase price of building land	taken	request	CSO_4
x730	Real estate	Share of unoccupied dwellings	taken	online	CSO_10
x800	Geomorphology	Mean altitude of municipality	generated	request	ArcData_1
x810	Geomorphology	Mountainousness of municipality	generated	request	ArcData_1
x820	Geomorphology	Mountainousness of municipality relative to its area	calculated	request	ArcData_1
x905	Climate	Annual average precipitation	generated	purchase	CHMI_12
x915	Climate	Annual average temperature	generated	purchase	CHMI_13

The last column of the Table 9 consist **Source tags**, which are the key for finding more information about the data sources of explanatory variables in the table below. The Table 10 provides the information about the organisation providing the data, name of the data source (data file) and if possible, also the URL for downloading the data [accessed: 09. 06. 2020]. For most of the data available only upon request, at least a preview of the source heading is provided. The names of the data sources and their headings were kept in the original language (Czech) to ease their search and possible communication with the data providers for those interested in obtaining the same sources.

Table 10: List of all data sources for the explanatory variables supplementing the Table 9

Source tag	Heading	Source details
ArcData_01	**Name of provider**	ARCDATA PRAHA, s.r.o.
	Source name	ArcCR500_v30.gdb
	URL/Preview	not available
CSO_02	**Name of provider**	Czech Statistical Office
	Source name	Průměrné ceny rodinných domů v ČR v letech 2010 - 2012 v závislosti na velikosti obcí a stupni opotřebení (v Kč/m3)

Source tag	Heading	Source details
	Source heading preview	Název kraje | Název okresu | Období 2010 - 2012 | VELIKOST OBCÍ (obyvatelé) | do 1999 | 2000 - 9999 | 10000 - 49999 | 50000 a více | Kupní cena
CSO_03	Name of provider	Czech Statistical Office
	Source name	Průměrné ceny bytů v ČR v letech 2010 - 2012 podle okresů v závislosti na velikosti obcí a stupni opotřebení (v Kč/m2)
	Source heading preview	Název kraje | Název okresu | Období 2010 - 2012 | VELIKOST OBCÍ (obyvatelé) | do 1999 | 2000 - 9999 | 10000 - 49999 | 50000 a více | Kupní cena
CSO_04	Name of provider	Czech Statistical Office
	Source name	Průměrné kupní ceny stavebních pozemků v ČR dle okresů a velikosti obcí v letech 2010 - 2012 (v Kč/m2)
	Source heading preview	Název okresu | Velikost obcí | Kupní ceny v letech | 2010 | 2011 | 2012
CSO_05	Name of provider	Czech Statistical Office
	Source name	prac_mista.xlsx
	Source heading preview	Kód obce | Obec místa obvyklého pobytu | Zaměstnaní | Dojíždějící | Vyjíždějící | Saldo dojížďky | Obsazená pracovní místa | Obsazená pracovní místa na 1 000 zaměstnaných obvykle bydlících v obci
CSO_06	Name of provider	Czech Statistical Office
	Source name	Tab. 112 Obyvatelstvo podle pohlaví a podle věku, rodinného stavu a nejvyššího ukončeného vzdělání v obci
	URL	https://vdb.czso.cz/vdbvo2/faces/cs/index.jsf?page=vystup-objekt-parametry&pvo=OTOB112&sp=A&pvokc=&katalog=30710&z=T
CSO_07	Name of provider	Czech Statistical Office
	Source name	Tab. 113 Obyvatelstvo podle pohlaví a podle ekonomické aktivity v obci
	URL	https://vdb.czso.cz/vdbvo2/faces/cs/index.jsf?page=vystup-objekt-parametry&pvo=OTOB113&sp=A&pvokc=&katalog=30713&z=T
CSO_08	Name of provider	Czech Statistical Office
	Source name	Tab. 114 Zaměstnaní podle pohlaví a podle odvětví ekonomické činnosti a podle nejvyššího ukončeného vzdělání v obci
	URL	https://vdb.czso.cz/vdbvo2/faces/cs/index.jsf?page=vystup-objekt-parametry&pvo=OTOB114&sp=A&pvokc=&katalog=30712&z=T

Source tag	Heading	Source details
CSO_09	**Name of provider**	Czech Statistical Office
	Source name	Tab. 116 Zaměstnaní podle postavení v zaměstnání a podle věku a pohlaví v obci
	URL	https://vdb.czso.cz/vdbvo2/faces/cs/index.jsf?page=vystup-objekt&pvo=OTOB116&z=T&f=TABULKA&katalog=30713&u=v185__VUZEMI__43__554979
CSO_10	**Name of provider**	Czech Statistical Office
	Source name	Tab. 118 Bytový fond v obci
	URL	https://vdb.czso.cz/vdbvo2/faces/cs/index.jsf?page=vystup-objekt&pvo=OTOB118&z=T&f=TABULKA&katalog=30731&u=v61__VUZEMI__43__554979
CSO_11	**Name of provider**	Czech Statistical Office
	Source name	Vybrané ukazatele pro územně analytické podklady za obec
	URL	https://vdb.czso.cz/vdbvo2/faces/cs/index.jsf?page=vystup-objekt&pvo=UAP01&z=T&f=TABULKA&katalog=31716&c=v94~3__RP2011&u=v4__VUZEMI__43__554979
CHMI_12	**Name of provider**	Czech Hydrometeorological Institute
	Source name	SRA_shape.shp
	URL/Preview	not available
CHMI_13	**Name of provider**	Czech Hydrometeorological Institute
	Source name	T_shape.shp
	URL/Preview	not available
MF_14	**Name of provider**	Ministry of Finance of the Czech Republic
	Source name	Podíl jednotlivých obcí na stanovených procentních částech celostátního hrubého výnosu daně z přidané hodnoty a daní z příjmů
	URL	https://www.mfcr.cz/cs/legislativa/legislativni-dokumenty/2012/vyhlaska-c-281-2012-sb-6529
MT_15	**Name of provider**	Ministry of Transport of the Czech Republic
	Source name	Druh pro jednotlivé obce
	URL	https://www.mdcr.cz/getattachment/Statistiky/Silnicni-doprava/Centralni-registr-vozidel/Statistika-2-2016-(k-1-7-2016)/Statistiky-vyplyvajici-z-Centralniho-registru-vozi/20-obec_dr.xls.aspx?lang=cs-CZ

The correlation matrix of municipality-related explanatory variables can be seen in annex [B], their box and whisker plots then in annex [C].

4.2.1.1 Commentary on calculated variables

This chapter names the inputs to the calculation and describes how the selected explanatory variables were calculated. Further notes are added if needed.

Age index (x140) – The Share of inhabitants with age 65+ (x130) divided by the Share of children aged 0-14 (x110)

Share of employees without high school diploma (x220) – The Share of employees with elementary education (x200) plus the Share of employees with apprenticeship certificate (x210)

Share of employees with high school diploma or higher (x270) – A sum of the following variables:

- Share of employees with high school diploma (x230)
- Share of employees with completed extension study (x240)
- Share of employees with tertiary technical education (x250)
- Share of employees with university degree (x260)

Average length of the education attendance of employees (x280) – A sum of the following products:

- Number of employees with elementary education times 8.5 year
- Number of employees with apprenticeship certificate times 11.5 year
- Number of employees with high school diploma times 12.5 year
- Number of employees with completed extension study times 14.5 year
- Number of employees with tertiary technical education times 14.5 year
- Number of employees with university degree times 18 year

divided by the total sum of:

- Number of employees with elementary education
- Number of employees with apprenticeship certificate
- Number of employees with high school diploma
- Number of employees with completed extension study
- Number of employees with tertiary technical education
- Number of employees with university degree

Ratio of economically inactive to active inhabitants (x340) – The Number of economically inactive inhabitants divided by the number of economically active inhabitants

Ratio of self-employed to economically active inhabitants (x350) – The Number of self-employed inhabitants divided by the number of economically active inhabitants

Ratio of occupied jobs to economically active inhabitants (x360) – The Number of occupied jobs divided by the number of economically active inhabitants

Ratio of inhabitants employed in key sectors (x370) – The Total number of employees in economic activity groups:

- Wholesale and retail; repair and maintenance of motor vehicles
- Transport and storage
- Accommodation, catering, hospitality
- Information and communication activities
- Finance and insurance
- Real estate activities
- Professional, scientific and technical activities
- Administrative and support service activities
- Public administration and defence; compulsory social security
- Education
- Health and social care

(These economic activity groups are representing the tertiary and partially also the quaternary and quinary sectors, thus the abbreviation 'T+')

divided by the total number of employees in economic activity groups:

- Agriculture, forestry, fishing
- Manufacturing industry
- Construction

(These economic activity groups are representing the key part, but not all of the economic activity groups of the primary and the secondary sectors, thus the abbreviation 'PS-').

The reason for using only partial sectors is that the source table (Tab. 114 Zaměstnaní podle pohlaví a podle odvětví ekonomické činnosti a podle nejvyššího ukončeného vzdělání v obci) does not include numbers of employees in the less frequent economic activity groups such as:

- Mining and quarrying
- Production and distribution of electricity, gas, heat and air conditioning
- Water supply; sewerage, waste management and remediation activities
- Cultural, entertainment and recreational activities
- Other activities

- Activities of households as employers; Undifferentiated goods- and services-producing activities of households for own use
- Activities of extraterritorial organizations and bodies

The reason for using aggregated sectors is that some of the smallest municipalities do not have any employee employed either in the partial primary sector or the tertiary sector, which hindrances the calculation. Thus, the partial primary and the partial secondary sectors were aggregated into PS- and the tertiary sector was aggregated with the partial quaternary and the partial quinary sector into T+.

Share of municipal inhabitant on national tax revenue (x400) – The Percentage share of the municipality in the tax revenue according to § 4 par. 1 let. b) to f) of the Act on Budgetary Determination of Taxes divided by the number of inhabitants of the municipality. The shares come from the attachment of the Decree No. 281/2012 Coll. with effect from 1 September 2012 issued by the Ministry of Finance of the Czech Republic. However, attachment is based on tax revenue from previous period, thus reflecting the time of the national population census.

Car ownership (x410) – The Number of passenger cars (personal automobiles) in the municipality divided by the number of inhabitants of the municipality times 1 000. The data on number of passenger cars is to date 08. 02. 2013 as this is the closest to the decisive moment on midnight from 25. 03. to 26. 03. 2011, which can be found in the archive of the Ministry of Transport.

Total number of workers, pupils, students and apprentices (x430) – The Number of employed inhabitants plus the total number of pupils, students and apprentices

Gross population density (x500) – The Number of inhabitants in the municipality divided by the total area of the municipality

Net population density (x510) – The Number of inhabitants in the municipality divided by the total size of built-up areas within the municipality

Mountainousness of municipality relative to its area (x820) – The Mountainousness of municipality (x810) divided by the total area of the municipality. See chapter 4.2.1.2 for details how the mountainousness of municipality was generated.

4.2.1.2 Commentary on generated variables

This chapter describes how the selected explanatory variables were generated. Further notes are added if needed.

Municipality organizing Mass City Transport (x600) – This is the only discrete binary explanatory variable used in this research. Value 1 was manually added to all municipalities (towns or cities), which are organising a municipal system of mass city transport on their premises. In the case of mass city transport systems or companies jointly belonging to multiple municipalities (e.g. the Transport company of the towns of Most and Litvínov), 1 was marked at all organising municipalities (thus at both towns). List of municipalities organising their own mass city transport was made based on a research of various Internet sources to verify the existence of the mass city transport systems in 2011. Value of 0 was recorded in the case of all other municipalities.

Distance to the closest train stop or station (x610) – This variable was generated in the GIS software ArcMap using the source database ArcCR500 version 3.0, relevant to 2011, kindly provided by the company ARCDATA PRAHA, s.r.o. The database includes a layer of singular points for all railway stops and stations in Czechia as well as layer of singular points for all municipalities in Czechia existing in March 2011. The layer of train stops and stations required modification to include only stops and stations used for passenger transport in March 2011 during work days, thus usable for daily commute to work or school at the time of national population census. Thus, archive version of the train tables from year 2010/2011 (VYKA, 2010) were used to verify if any passenger trains were stopping at the stops and station there during weekdays in March. Redundant stops and stations were removed from the layer. A function 'Near' of ArcMap was then used to find the closest stop or station to each municipality and record the mutual distance. The distance is an air distance between the point of stop/station and the point of municipality. Positions of the points were not edited. The last step was to export the table of municipalities with the distance to the closest train stops or stations and name of the respective stop or station.

Mean altitude of municipality (x800) – This variable was calculated as an average of the pseudo-lowest and the pseudo-highest altitude in the municipality – see x810 below for description.

Mountainousness of municipality (x810) – This variable was generated in the GIS software ArcMap using the source database ArcCR500 version 3.0, relevant to 2011, kindly provided by the company ARCDATA PRAHA, s.r.o. The database includes a layer of 5-metre interval contour lines and a layer of polygons of all municipalities in Czechia as existed in March 2011. The polygons were representing the total area of the municipalities. A function 'Intersect' of ArcMap was used to find the lowest and the highest contour line intersecting the municipality polygons, thus finding the pseudo-lowest altitude within the municipality limits and the pseudo-highest altitude in the municipality. It was necessary to find the lowest and the highest altitude manually in the case of municipalities, which were not intersect by any or by only one contour line. A manual search was done using a free map portal

www.mapy.cz. It was also necessary to manually correct the pseudo-lowest altitude in municipalities, which have deep brown coal mine (open pit mine). The pseudo-lowest altitude of such municipalities was set to altitude of the upper edge of the mine pit. Once each municipality had the pseudo-lowest altitude and the pseudo-highest altitude, these were exported as a table. The mountainousness was then calculated as difference between the pseudo-lowest altitude and the pseudo-highest altitude. The greater the difference between these two, the greater the mountainousness of municipality is.

Annual average precipitation (x905) – This variable was generated in the GIS software ArcMap using the source database ArcCR500 version 3.0, relevant to 2011, kindly provided by the company ARCDATA PRAHA, s.r.o. and the source shape file SRA_shape.shp purchased from the Czech Hydrometeorological Institute (CHMI). The precipitation shape file includes polygons of 10 different precipitation zones with annual average precipitation for the period 1991-2010. The precipitation zones in Czechia can be seen in Figure 16.

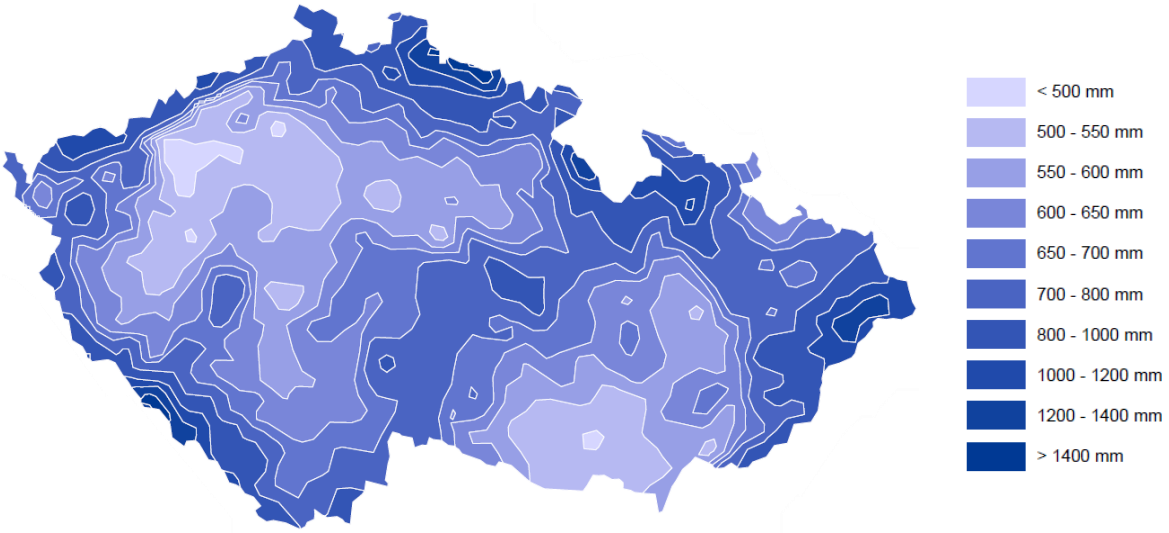


Figure 16: Annual average precipitation zones in Czechia for the period 1991-2010, source CHMI

The polygons of the precipitation zones were then intersected with the polygons of the municipalities. Each municipality was assigned to a precipitation zone, which was covering the majority of its total area. The resulting prevailing precipitation zones in every municipality were then exported in form of a table, where the names of the precipitation zones (the precipitation intervals) were replaced by codes 1 through 10 as shown in Table 11.

Table 11: Conversion table of Annual average precipitation zones to codes

Precipitation Interval	Code
< 500 mm	1
500 - 550 mm	2
550 - 600 mm	3
600 - 650 mm	4

Precipitation Interval	Code
650 - 700 mm	5
700 - 800 mm	6
800 - 1000 mm	7
1000 - 1200 mm	8
1200 - 1400 mm	9
> 1400 mm	10

Annual average temperature (x915) – This variable was generated in the GIS software ArcMap using the source database ArcCR500 version 3.0, relevant to 2011, kindly provided by the company ARCDATA PRAHA, s.r.o. and the source shape file T_shape.shp purchased from the Czech Hydrometeorological Institute (CHMI). The temperature shape file includes polygons of 8 different temperature zones with annual average temperature for the period 1991-2010. The temperature zones in Czechia can be seen in Figure 17.

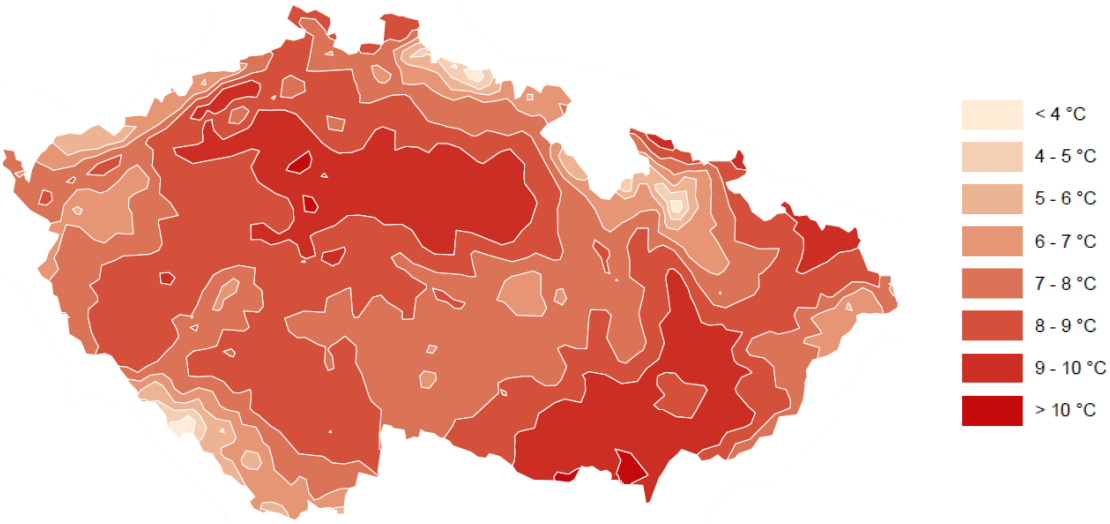


Figure 17: Annual average temperature zones in Czechia for the period 1991-2010, source CHMI

Similarly to precipitation zones, each municipality was then assigned to a temperature zone, which was covering the majority of its total area. The resulting prevailing temperature zones in every municipality were then exported to a table, where the names of the temperature zones (the temperature intervals) were replaced by codes 1 through 8 as shown in Table 12.

Table 12: Conversion table of Annual average temperature zones to codes

Temperature interval	Code
< 4 °C	1
4 - 5 °C	2
5 - 6 °C	3
6 - 7 °C	4
7 - 8 °C	5

Temperature interval	Code
8 - 9 °C	6
9 - 10 °C	7
> 10 °C	8

4.2.1.3 Commentary on additional variables

Since the decision tree regression can process mutually dependent explanatory variables, there was no restriction to add more variables from the same group if it was easy to do or if influence to modal split was expected.

Share of inhabitants aged 15-64 (x120) – variable simply taken from the source completing the range of age groups in the data set of explanatory variables

Employee education explanatory variables x210, x230, x240 and x250 – variables simply taken from the source completing the range of employee education groups in the data set of explanatory variables

Share of employees without high school diploma (x220) and **Share of employees with high school diploma or higher (x270)** – variables simply calculated from the source data. The reason for adding these aggregated variables is to research, whether the high school diploma is the breaking point, which has the influence to the local travel behaviour

Municipality organizing Mass City Transport (x600) – variable generated because it is expected to have positive association with the use of the transport mode Mass city transport

Mountainousness of municipality variables x810 and x820 – variables generated because these are expected to have negative association with the use of the transport mode Bike

4.2.1.4 Commentary on omitted variables

Some of the explanatory variables included in the review were not included in the final set of explanatory variables, typically due to absence of relevant data, eventually due to redundancy.

Gross domestic product per capita and/or **Volume index of industrial output** – such data are not available on the municipal level

Average income of inhabitant/voter, Average wage and **Number of inhabitants with an income above average** – data not available on the municipal level

Tax base of a median voter – the median voter is not defined in the Czech statistics system

Tax base of self-employed inhabitants – data not available on the municipal level

Annual average number of working people – data not available on the municipal level

Ratio of number of job vacancies to number of economically active inhabitants – data not available on the municipal level

Share of independent farmers per 1 000 inhabitants – data not available on the municipal level

Educational index – this variable was assessed redundant since the variable Average length of the education attendance of employees (x280) is based on similar, but more detailed data inputs

Altitude of centre of settlement – similar variable Mean altitude of municipality (x800) was prioritized over this one due to much easier generation

Mean slope – data not publicly available

Variables on accessibility to transport hubs and networks – only the locations of train stops and stations were available

Presence of congestion – data not publicly available

Price of monthly ticket for the public transport – this variable makes sense for travelling inside of a city or area, where a traveller can freely move for the one price paid on a time basis. The local intermunicipal commute is not the case as the price vary according to the travelled distance

Annual average number of rainy days – similar variable Annual average precipitation (x905) was prioritized over this one due to available data

4.2.2 Variables related to connections of municipalities

Each O-D pair of two municipalities can be (and mostly is) connected by multiple transport **connections**. It is because multiple transport modes can be used for journeys in one O-D pair. Different modes then use various routes and timeslots, which further diversifies the set of options how to move in time and space between the municipalities of origin and destination. Different criteria were used for distinguishing the connections made by public transport and individual transport modes.

The connections using the **single public transport modes Bus, Train and MCT** were distinguished based on the used transport mode and timeslot. The municipalities of origin and destination are assumed to be connected daily by multiple connections of each mode. To find the frequency of connections by each single public transport mode, following conditions apply:

- the frequency of the public transport connections is represented by a number of connections per day, which is given by the number of unique arrivals to the municipality of destination found in the publicly available timetables;
- in the case of multiple connections arriving at the same time, only the connection with the shortest travel time was counted;
- in the rare case, when one connection (C1) departs after the other (C2) but is faster and manages to arrive before the C2, then only the connection C1 is counted;
- both direct connections and connections requiring transfer were counted if the above conditions were fulfilled.

The connections using the **combined public transport modes PTplus and PTcom** were not distinguished in any way, because the transfer points between the modes and the order in which the modes were used, were not recorded by the national population census. Thus, it is not possible to deduce, for which part of the journey, the public transport was used, which makes it impossible to find any further details about the connections in the timetables or the maps.

The connections made by the **individual transport modes** (Driver, CarPass, Bike, Walk) were distinguished based only on the used transport mode. There is no point of further distinguishing them based on the timeslots as there are no given timeslot nor timetables for the individual transport. The times of departure and arrival were not recorded by the national population census. Nor were recorded the details about the used routes. Thus, an assumption on the most likely route was made for each individual transport mode based on the publicly available maps. Within the above provided framework, the connections made by private transport within one O-D pair were defined as follows:

- **Driver** – the municipalities of origin and destination are assumed to be connected by one Driver type connection, which is routed on the shortest road connection of the two municipalities;
- **CarPass** – only one CarPass type connection analogical to Driver type connection;
- **Bike** – the municipalities of origin and destination are assumed to be connected by one Bike type connection, the bike connection is assumed to be routed on the shortest road connection of the two municipalities if no cycle path is available between them or on the available cycle paths, if the routing via cycle paths does not significantly prolong the travelled distance;
- **Walk** – only one Walk type connection analogic to Bike type connection, only employing pedestrian paths and combined pedestrian and cycling paths, if possible.

All the journeys are taken using one of the possible connections between the municipalities of origin and destination. The connections are then a framework or an envelope for the journeys taken by the

respective transport modes within specified timeslots. Therefore, the **journey characteristic indicators** reviewed in chapter 2.1.3. describing the journeys can be also used for describing the characteristics of the connections. Therefore, the explanatory variables related to connections between municipalities will be based on the journey characteristic indicators such as travel distance, travel time and weather conditions.

However, some of the **spatial indicators** will be used as the connection-related explanatory variables, typically the availability and proximity of private transport networks, the availability and proximity of public transport hubs, and frequency of the public transport connections. The private transport networks are primarily the public roads, cycle paths and pedestrian paths, eventually the combined pedestrian and cycling paths.

Most of the characteristics related to the connections between the municipalities must have been manually extracted from the publicly available sources one by one. Due to such enormous data collection burden, the characteristics were collected only for a **limited sample of O-D pairs**. The sample includes the O-D pairs related to an administrative district of one municipality with extended powers (MEP), which is reported as a natural centre of daily commute catchment area. Only O-D pairs having either origin or destination in the administrative district were included into the sample along with the ones having there both origin and destination.

The selected administrative district of MEP is **Šternberk** in the Olomouc region. The reasons for choosing Šternberk are as follows:

- the administrative district includes both flat land (Upper Moravian valley) and mountainous terrain (Low Jeseník Mountains);
- the administrative district includes areas with both high and low population density;
- sufficient number of independent municipalities within the administrative district – 21 municipalities with all together 23 288 inhabitants (by the time of population census 2011);
- the capital of the administrative district – the Town of Šternberk – is a mid-size MEP with 13 574 inhabitants (2011);
- each of the two regional railway lines passing through the administrative district has a significantly different daily frequency of trains;
- the author can rely on an excellent knowledge of the local environment.

The final list of all connection-related explanatory variables used in the connection-related models of modal split dependency can be seen in Table 13. The list contains **33 explanatory variables**. The table distinguishes different demands on data pre-processing and availability, already described in chapter 4.2.1 on municipality-related explanatory variables. Specific details on data generation and calculation

are described in chapters 4.2.2.1 and 4.2.2.2. Regarding the data availability, some data sources were available only upon request, some even had to be officially purchased, but the majority of the data are available online, unfortunately not in machine-readable format.

The **green colouring** marks the variables, which were added to the set of explanatory variables in comparison to the Table 1 in chapter 3.7.1. The variables, which were omitted in comparison to the Table 1 are obviously not recorded in Table 13, but will be commented in chapter 4.2.2.3.

Table 13: List of all connection-related explanatory variables used in the connection-related models (C)

Code	Group	Explanatory variable	Preparation	Available	Source tag
x10	transport	Distance to Bus stop	generated	online	MAPY_01
x11	transport	Distance from Bus stop	generated	online	MAPY_01
x12	transport	Total distance to and from Bus stops	calculated	online	MAPY_01
x13	transport	Number of connections by Bus	generated	online	IDOS_02
x20	transport	Distance to Train stop or station	generated	online	MAPY_01
x21	transport	Distance from Train stop or station	generated	online	MAPY_01
x22	transport	Total distance to and from Train stops or stations	calculated	online	MAPY_01
x23	transport	Number of connections by Train	generated	online	VYKA_03
x31	transport	Distance by I. class road	generated	online	MAPY_01
x32	transport	Distance by II. or III. class road	generated	online	MAPY_01
x34	transport	Proportion of distance by I. class road	calculated	online	MAPY_01
x40	transport	Shortcut / prolonging by Bike	calculated	online	MAPY_01
x41	transport	Shortcut / prolonging by Bike relative	calculated	online	MAPY_01
x42	transport	Distance by rural cycle path	generated	online	MAPY_01
x43	transport	Distance by rural cycle path relative	calculated	online	MAPY_01
x44	transport	Total elevation of Bike route	generated	online	MAPY_01
x45	transport	Total elevation of Bike route relative	calculated	online	MAPY_01
x60	travel distance	Air distance	generated	online	MAPY_01
x61	travel distance	Total distance by roads	generated	online	MAPY_01
x62	travel distance	Total distance by Bike	generated	online	MAPY_01
x63	travel distance	Level of neighbouring	generated	online	MAPY_01
x70	travel time	Travel time by Bus	calculated	request	CSO_11
x71	travel time	Travel time by Train	calculated	request	CSO_11
x73	travel time	Travel time by Driver	calculated	request	CSO_11
x74	travel time	Travel time by CarPass	calculated	request	CSO_11
x76	travel time	Travel time by Bike	calculated	request	CSO_11
x80	travel time	Travel time share of Bus	calculated	request	CSO_11
x81	travel time	Travel time share of Train	calculated	request	CSO_11
x83	travel time	Travel time share of Driver	calculated	request	CSO_11
x84	travel time	Travel time share of CarPass	calculated	request	CSO_11
x86	travel time	Travel time share of Bike	calculated	request	CSO_11
x90	climate	Averaged annual average precipitation	calculated	purchase	CHMI_12

Code	Group	Explanatory variable	Preparation	Available	Source tag
x91	climate	Averaged annual average temperature	calculated	purchase	CHMI_13

The last column of the Table 13 consist **Source tags**, which are the key for finding more information about the data sources of explanatory variables in the table below. The Table 14 provides the information about the organisation providing the data, name of the data source (data file) and if possible, also the URL for downloading the data [accessed: 15. 06. 2020]. For the data available only upon a request, at least a preview of the source is provided. The names of the data sources and their headings were kept in the original language (Czech) to ease their search and possible communication with the data providers for those interested in obtaining the same sources.

Table 14: List of all data sources for the connection-related explanatory variables supplementing the Table 13

Source tag	Heading	Source details
MAPY_01	**Name of provider**	Seznam.cz, a.s.
	Source name	mapy.cz
	URL	https://mapy.cz
IDOS_02	**Name of provider**	CHAPS s.r.o.
	Source name	idos.cz
	URL	https://idos.idnes.cz
VYKA_03	**Name of provider**	Miroslav Vyka
	Source name	Železniční jízdní řád 2010/2011
	URL	http://www.jizdni-rady.nanadrazi.cz/
CSO_11	**Name of provider**	Czech Statistical Office
	Source name	vyj_doj_mimoobecni_denni_vystup.xlsx
	Preview	see Figure 8 on page 65
CHMI_12	**Name of provider**	Czech Hydrometeorological Institute
	Source name	SRA_shape.shp
	URL/Preview	not available
CHMI_13	**Name of provider**	Czech Hydrometeorological Institute
	Source name	T_shape.shp
	URL/Preview	not available

The correlation matrix of connection-related explanatory variables can be seen in annex [B], their box and whisker plots then in annex [C].

4.2.2.1 Commentary on generated variables

This chapter describes how the selected explanatory variables were generated. Further notes are added if needed. The explanatory variables are described in a logical order, not in order of appearance in Table 13.

Distance to Bus stop (x10) and **Distance to Train stop or station (x20)** – These 2 explanatory variables were both generated the same way using the Planning tool at the map portal 'mapy.cz', which can provide distance measurements of routes on transport infrastructure. As a Start point was entered the name of the municipality of origin. As an End point was entered the location of the stop or station selected in the map. The transport mode was selected 'By foot' and its setting to 'Short'. The result was the shortest walking distance from the centre of municipality of origin to the respective stop or station. The centre of the municipality is predefined by the portal. The distance was mostly measured to stop or station closest to the centre of municipality. Exceptionally, the stop or station providing higher frequency of public transport connections was selected, if it did not cause a significant prolonging of the walking distance.

Distance from Bus stop (x11) and **Distance from Train stop or station (x21)** – These 2 explanatory variables were both generated analogically to the previous two, only with an opposite selection of Start and End points in the Planning tool.

Number of connections by Bus (x13) – Number of connections was counted using 'Bus' timetables at the public transport timetables portal 'idos.cz'. In idos.cz, it is not possible to ask questions on public transport connection regarding the past, thus it was not possible to ask for the connections available at the time of the national population census (March 2011). A search for an archive timetables for all Bus connection in the area was not successful either (unlike the Train timetables). Thus, present Bus connection were searched in idos.cz, neglecting the changes that could happen to Bus services between 2011 and 2020. The search in idos.cz was done in February 2020, thus in similar season as the last national population census. Luckily, it was done before the state of emergency was declared in Czechia in mid-March 2020 due to COVID-19, which has affected the public transport services.

Number of connections by Train (x23) – Number of connections was counted using archive versions of Train timetables 2010-2011, which were valid at the time of the last national population census.

Distance by I. class road (x31) and **Distance by II. or III. class road (x32)** – These 2 explanatory variables were both generated the same way using the Planning tool at the map portal 'mapy.cz'. As a Start point was selected the point, where the car route entered the I., II. or III. class road and the End point was the point, where the car route left the road of the respective class. The transport mode was

selected ‘By car’ and its setting to ‘Short’. Distances by II. and III. class roads were aggregated. This is because the II. and III. class roads are owned and managed by the self-governing regions, while the I. class roads by the national government (Road and motorway directorate of the Czech Republic). Moreover, distinguishing the II. and III. class roads would disaggregate those variables too much.

Air distance (x60) – This explanatory variable was generated using the Planning tool at the map portal ‘mapy.cz’. As a Start point was entered the municipality of origin. As an End point was entered the municipality of destination. The Planning tool then pointed out the centres of both municipalities used in the mapy.cz geographical information system. The air distance between these two points was then measured using the tool ‘Distance and Area measurement’.

Total distance by roads (x61) – This explanatory variable was generated using the Planning tool at the map portal ‘mapy.cz’. As a Start point was entered the municipality of origin. As an End point was entered the municipality of destination. The transport mode was selected ‘By car’ and its setting to ‘Short’. The result was the shortest by road distance from the centre of municipality of origin to the centre of municipality of destination.

Note: The real-world journeys of the commuters are of course not always only between the centres of the municipalities, but between different places within them. However, the exact routes or specific point of origin and destinations of the journeys were not provided, due to personal data protection, as these could be the home or work addresses of the citizens. Thus, the assumed origin and destination points in the centres of municipalities represent pseudo average location of these points within the urban areas of the municipalities. The possible bias caused by this assumption is partially limited by smaller size of municipalities in the research of local commute, from which the large towns and cities were removed.

Total distance by Bike (x62) – This explanatory variable was generated using the Planning tool at the map portal ‘mapy.cz’. As a Start point was entered the municipality of origin. As an End point was entered the municipality of destination. The transport mode was selected ‘By bike’. Both settings ‘MTB’ (Mountain bike) and ‘Road bike’ were tried out, to find out which is resulting in a shorter distance and this, was then recorded. The Bike route was further manually adjusted if needed to use cycle paths, if such use did not cause significant prolonging of the travelled distance. The result was the shortest distance from the centre of municipality of origin to the centre of municipality of destination ridable by Bike with a preferences of cycle path use.

Total elevation of Bike route (x44) – During gathering of the variable ‘Total distance by Bike’, the altitude profile of the route was displayed in the feature of the same name, provided as a part of the

Planning tool at the map portal 'mapy.cz'. In addition to the profile itself, the feature also displays two values for total accent and total descent of the route. These two values were added to obtain the Total elevation of Bike route.

Distance by rural cycle path (x42) – The distance travelled by rural cycle path within the Bike route was measured using the Planning tool at the map portal 'mapy.cz'. As the Start/End point were entered the points in a map, where the cycle path in rural area starts or where the character of the cycle path changes from urban to rural. Only those cycle paths were measured, whose entire rural section between any two municipalities were completed before the time of national population census (March 2011). The reasons for measuring only the rural sections of the cycle paths were as follows:

- The average speed difference of Bike traffic and automobile traffic is much higher on the rural roads with speed limit of 90 km/h than in the case of the urban roads with speed limit of 50 km/h, considering the standard speed limits of Czechia. A higher speed difference is expected to be associated with a higher road safety risk to the cyclists.
- There is generally only one cycle path between two municipalities in the O-D pair and its rural part is thus the only (unavoidable) part of the Bike route between the two municipalities, which the cyclists most likely use. On the other hand, within the urban areas of the municipalities, the travellers can take more routes to get from their point of origin to the beginning of the rural cycle path. Thus, their route within the urban area is uncertain.

Level of neighbouring (x63) – This variable describes how close neighbours are the two municipalities in the O-D pair. The idea behind this variable is based on the hypothesis that not only the increasing distance of the municipalities could have influence to the modal split, but also the increasing number of transport obstacles or complications on the way between them. Example of such complication could be another municipalities or settlements and busy roads in them. The number of such complications on the way would then increase with the increasing number of settlements on the way. This might be especially influential in the case of the active travel modes (Bike, Walk) used for school commute. The parents might rather let their children ride a bike to 5 km distant neighbouring municipality located across the field from the municipality of origin, than to 3 km distant municipality, to which the road is passing through another settlement, where a busy road needs to be crossed.

However, there is **no clear-cut solution how to** assess the level of neighbouring of municipalities. The reason for that is given by the complicated governmental structure of Czechia. The shapes of the areas representing the municipality limits are commonly far from the perfect hexagons, which would allow to assess the neighbouring easily like in the honeycomb structure. On the contrary, the shapes can be as complicated as a shape of an area of the Town of Šternberk, which can be seen in Figure 18.

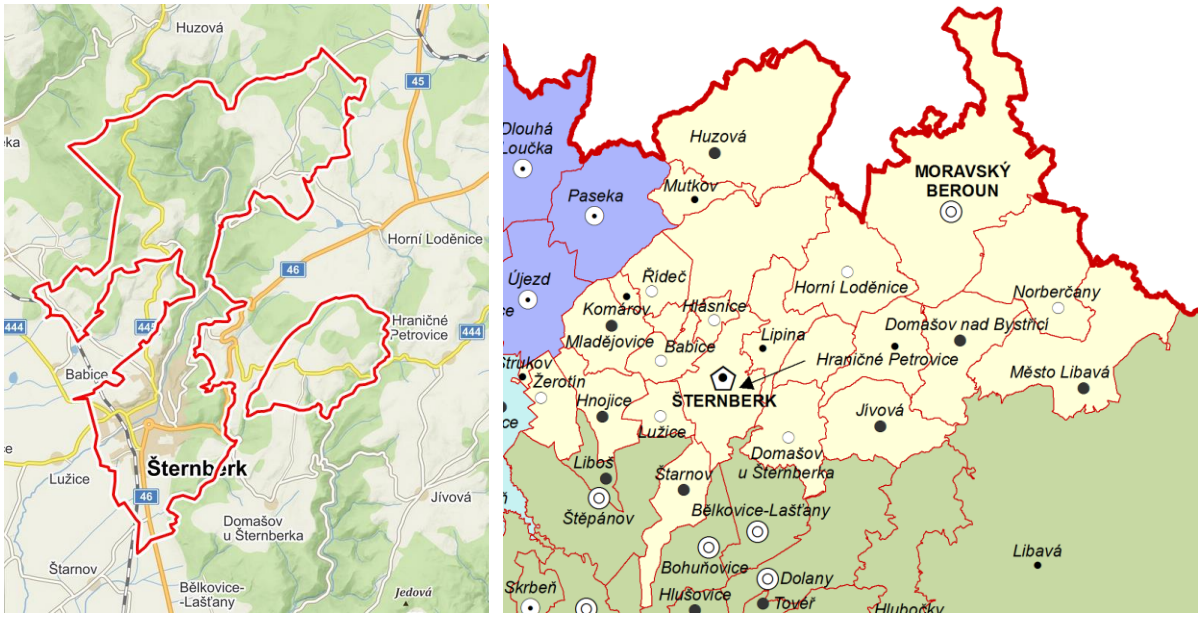


Figure 18: Presentation of the complicated municipal structure on the example of the Town of Šternberk, picture on the left is taken from the map portal mapy.cz, picture on the right is from CSO (2016)

Moreover, not all the settlements or villages are an independent self-governing municipality. The Figure 19 is showing that the complicated shape of the administrative area of the Town of Šternberk includes another six settlements in the surroundings of the actual town. From the practical point, the settlements, which are geographically neighbouring the core urban area of the Town of Šternberk are not always an administrative neighbour and vice versa. Thus, it is clear that sharing the border does not make two municipalities close neighbours and the level of neighbouring **cannot be assessed based on sharing the municipal border**.

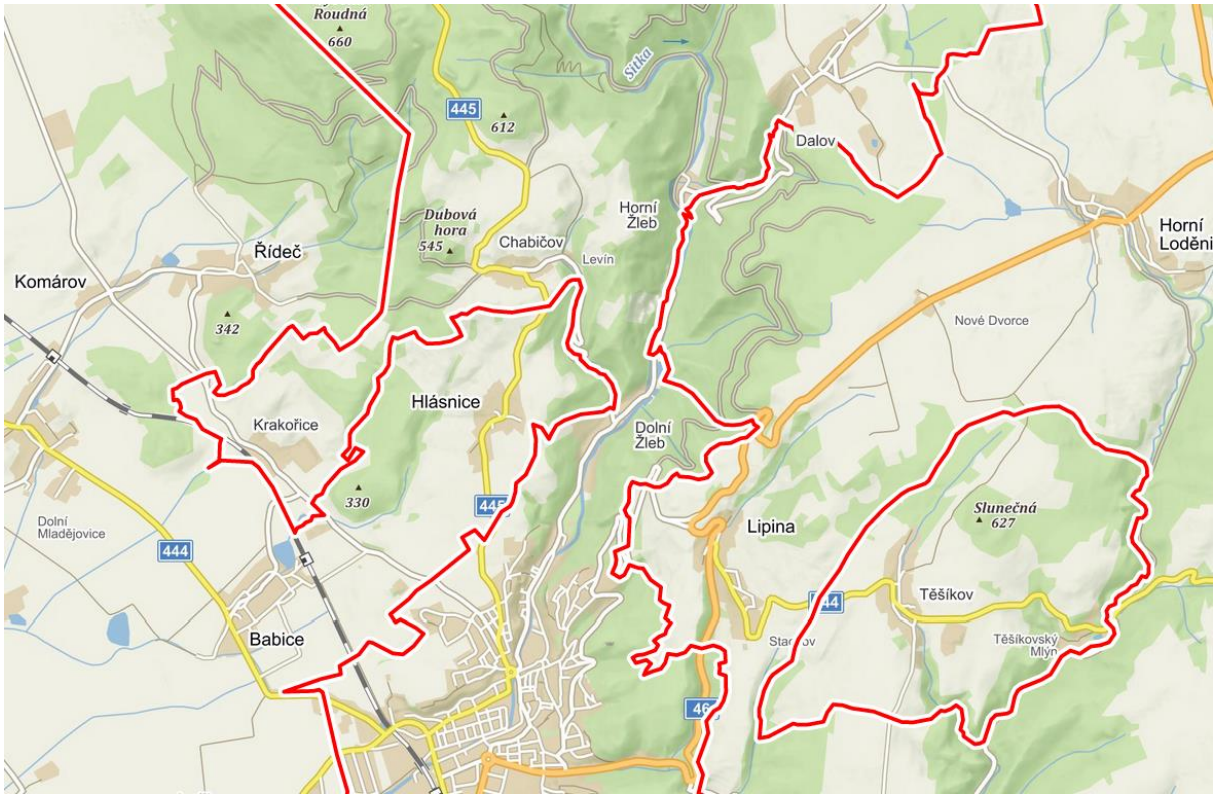


Figure 19: Local settlements Těšíkov, Krakořice, Dolní Žleb, Horní Žleb, Chabičov and Dalov, which are not independent municipalities but a part of the Town of Šternberk, source: mapy.cz

In the end, the level of neighbouring of municipalities was generated **based on the number of settlements on the route** from one municipality to another. At first, the route from municipality of origin to municipality of destination was generated using the Planning tool at the map portal 'mapy.cz'. The generated route was then studied in the map and each settlement on the route was counted regardless of its status (no matter if it was an independent municipality or not). If no settlement was on the route, the level of neighbouring was 0. One settlement on the route means level of neighbouring equal 1, two settlements means level 2 and three or more counted settlements results in level of neighbouring equal 3. Thus, the level of neighbouring is one of the few discrete explanatory variables with only four possible values 0, 1, 2 and 3.

4.2.2.2 Commentary on calculated variables

This chapter names the inputs to the calculation and describes how the selected explanatory variables were calculated. Further notes are added if needed.

Total distance to and from Bus stops (x12) – The Distance to Bus stop (x10) plus the Distance from Bus stop (x11).

Total distance to and from Train stops or stations (x22) – The Distance to Train stop or station (x20) the plus Distance from Train stop or station (x21).

Proportion of distance by I. class road (x34) – The Distance by I. class road (x31) divided by the Total distance by roads (x61).

Shortcut / prolonging by Bike (x40) – The Total distance by roads (x61) minus the Total distance by Bike (x62).

Shortcut / prolonging by Bike relative (x41) – The Shortcut / prolonging by Bike (x40) divided by the Total distance by roads (x61). The Shortcut / prolonging by Bike is then relative to the Total distance by roads, which is a benchmark for realistic mutual distance of municipalities.

Distance by rural cycle path relative (x43) – The Distance by rural cycle path (x42) divided by the 'Total rural length of the Bike route'. The total rural length of the Bike route was measured using the Planning tool at the map portal 'mapy.cz'. The total rural length of the Bike route could have been composed of multiple sections going through rural areas (outside of urban areas). As the Start/End points were entered the points in map representing the change from urban to rural environment and vice versa. The total rural length of the Bike route was measured only for the O-D pairs with non-zero value of the variable Distance by rural cycle path (x42), thus it is not available for all O-D pair and it is not used as an explanatory variable on its own.

Total elevation of Bike route relative (x45) – The Total elevation of Bike route (x44) divided by the Total distance by Bike (x62). The Total elevation of Bike route is then relative to the Total distance by Bike.

Travel time by Bus (x70) – The Travel time in minutes averaged over all journeys taken by Bus on the respective O-D pair, see chapters 4.1.4 and 4.1.5 for more details. The explanatory variables **x71**, **x73**, **x74** and **x76** were calculated analogically.

Travel time share of Bus (x80) – The 'Travel time share' of transport mode is a newly introduced explanatory variable. The idea behind it is to make an analogy to the modal share. The Travel time share (TTS) is calculated according to one of the three following simple formulas. Boundary conditions deciding on which formula to use are based on a number of transport modes used (n) within the O-D pair and if the mode in question was used within the O-D pair – if it has any average travel time.

$$TTS_i = 1 - \frac{ATT_i}{SATT}, \qquad for\ n\ > 1\ and\ ATT_i\ > 0 \tag{20}$$

$$TTS_i = 1, \qquad for\ n = 1 \tag{21}$$

$$TTS_i = 0, \qquad for\ ATT_i = 0 \tag{22}$$

where:

$TTS_i =$ travel time share of transport mode i within the O-D pair; $TTS_i \in \langle 0; 1 \rangle$; $i \in \langle 1; n \rangle$

$ATT_i =$ average travel time of transport mode i (an arithmetic mean of travel times on all journeys within the O-D pair, which were taken using the transport mode i)

$SATT =$ sum of average travel times of all transport modes used within the O-D pair

$n =$ number of transport modes used within the O-D pair

$$SATT = \sum_{1}^{n} ATT_i \tag{23}$$

Higher is the travel time share of a transport mode **the better** for the transport mode. The transport mode can achieve higher values of travel time share either by having shorter average travel time (ATT) than the competitors within the O-D pair or by having more of competing transport modes there, which will together contribute to larger sum of average travel times ($SATT$). The travel time share should represent how competitive is the average travel time of the transport mode in comparison with average travel times of the other modes within the same O-D pair. The higher is the travel time share of a transport mode the higher is the competitiveness of its average travel time.

The 'trouble' comes with the **O-D pairs where only one transport mode was used**. If using the formula (20) for calculation of the travel time share of the only transport mode used on the O-D pair, the result would be equal 0. This is in direct conflict with the stipulated definition that the higher the travel time share of a transport mode the better. The only transport mode used within the O-D pair (a singular transport mode) can be perceived as so competitive, it has beaten all the other transport modes and won a monopoly on transport services within the O-D pair. The singular transport mode should have the maximum travel time share, which is 1. That is assured by using the formula (21). It is important to note that the travel time share can be calculated only for modes with any average travel time. Modes without any travel time have travel time share equal 0 according to formula (22). In the end, the calculated travel time share values must be normalised to 1 in each O-D pair based on the number of modes used within the respective O-D pair.

The explanatory variables **x81**, **x83**, **x84** and **x86** were calculated analogically.

Averaged annual average precipitation (x90) – This variable is calculated as an average of the Annual average precipitation (x905) related to the municipality of origin and the Annual average precipitation (x905) related to the municipality of destination. It is then the average of two annual average precipitation values. Since the administrative district of MEP Šternberk includes both lowlands and

mountains, the precipitation in the origin and destination can vary even by 4 zones out of 10, see chapter 4.2.1.2 for more details.

Averaged annual average temperature (x91) – This variable is calculated analogically to the previous one. The temperature in the origin and destination can vary even by 3 zones out of 8.

4.2.2.3 Commentary on omitted variables

Some explanatory variables included in the review were not included in the final set of explanatory variables, typically because the necessary details were not collected during the national population census, eventually due to redundancy.

Travel cost – the travel cost is highly interrelated with the travel distance, which is much easier to gather, thus the travel distance was prioritised over the travel cost

Departure time – this information was not collected during the national population census

Trip chaining – this information was not collected during the national population census

Information – no records about information available to travellers before and during their commute were collected during the national population census

Interchange – this information was not collected during the national population census

4.3 Modal split dependency modelling

Three types of models were employed in the research. The first two are municipality-related and the third one is related to connections between the municipalities. The **type A model** is modelling the dependency of the transport mode shares on the explanatory variables related to the municipality of origin – the starting point A of the commuting journey in the O-D pair. The input data sets prepared for the type A models have the structure illustrated in Table 15. Analogically, the **type B model** is modelling the dependency of the modal shares on the characteristics of the municipality of destination – the end point B of the commuting journeys in the O-D pair. The type B model input data structure is illustrated in Table 16. Finally yet importantly, the **type C model** is modelling the dependency of the modal shares on the characteristics of the connections between the municipalities of the O-D pair. The type C model input data structure is illustrated in Table 17.

Table 15: Structure of the input data for the type A model

Municipality	**Municipality**	**Up to 39 explanatory variables**	**Share of transport mode *i***
A (origin)	B (destination)	X data relevant to municipality **A**	Y data of the O-D pair A→B

Table 16: Structure of the input data for the type B model

Municipality	**Municipality**	**Up to 39 explanatory variables**	**Share of transport mode *i***
A (origin)	**B** (destination)	X data relevant to municipality **B**	Y data of the O-D pair A→B

Table 17: Structure of the input data for the type C model

Municipality	**Municipality**	**Up to 33 explanatory variables**	**Share of transport mode *i***
A (origin)	B (destination)	X data relevant to connections (**C**) between municipalities A and B	Y data of the O-D pair A→B

The dependency of **modal share of each transport mode was modelled independently**. The transport mode 'Rest' aggregating the journeys taken by negligible modes was not modelled. Thus, each version of the model was used 9 times for the remaining 9 aggregated transport modes. Such 9 models are comprising a basic set of models, which are further evaluated together.

About the **number of models to be done**. Each set of consist of 9 models. The basic 3 types of models are the above described A, B and C. That would make 27 models. However, each of those types can be modelled using either filtered data set of O-D pairs or nonfiltered data set, thus increasing the number of required models to 54. Modelling independently the school and work commute will than again double the number of models to 108. Comparing the data sets of school and work commute, the work commute data set of O-D pairs is the larger one, thus more demanding for computing resources. Moreover, the nonfiltered data set is much larger and further increases the computation time. Also, the full set of explanatory variables is increasing the computational complexity of the model.

Thus, **it was decided to rationalise** the number of models to be done as well as data inputted. The nonfiltered data set of O-D pairs will be modelled only by using the linear regression models, which are less demanding in terms of the resources. The assumption is that the models made of nonfiltered data set will result in dependencies with weaker strength. In the case this assumption is verified, the nonfiltered data set will not be used any further.

The **type B model** has its own logic of selection of municipality-related explanatory variables into the model. It is based on the assumption that if a traveller from the municipality of origin (municipality A) decides on a modal choice for his/her travel to municipality of destination (municipality B), the age or education characteristics of the municipality B does not play any role in his/her decision-making process whatsoever. Thus, it is possible to avoid using the age and education relevant explanatory variables from the data set.

The modelling **will start with the** less demanding **school commute** so the proposed rationalisation methodology can be verified on a set of models taking less time to compute.

4.3.1 Decision tree modelling

The dependency of the modal split (the explained variable) on the explanatory variables was researched using decision tree models, theoretically described in chapter 2.2.1. The decision tree model provides a value of importance of each explanatory variable involved in explaining the dependency of modal split.

The models used in this research were built using the **Python programming language**. The modelling script is a self-customized version of a script provided by the scikit-learn community, described in work of Pedregosa et al. (2011), based on optimised version of the CART algorithm. The customizations were mainly related to import of input data, simulation of pruning methods, implementation of grid search function and exports of model outputs. The preview of the used script is in the annex [F]. Scikit-learn is an open source machine-learning library in Python. It is developed by a world-wide community, including experts on statistics, algorithms and software production. The quality of scikit-learn, its algorithms and documentation are universally acclaimed. The scikit-learn package is an example of object-oriented programming. Scikit-learn gives an opportunity to use its pre-developed and tested functions (objects), which execute series of commands performing the mathematical operations, for example the ones described in the theoretical chapter on Decision tree regression.

To be uploaded to the decision tree script in Python, the input data were saved as .csv file format. The first key task of the script is to split the uploaded data into the **test set** and **training set** of data. This split was done randomly in ratio 80 % training data and 20 % test data. No data were set aside for the validation data set, because the script includes cross-validation function, executing the multiple fold validation of the model on training data. To allow for repeatability of the modelling computation, the data split is set to have a constant random seed.

The function 'DecisionTreeRegressor' is responsible for the calculation of the decision tree model, having various **parameters** controlling the calculation. The full list of the parameters of this function and their default values can be seen in its documentation (DecisionTreeRegressor). The only three parameters which were actively set differently from the default values were the following:

max_depth – The maximum depth of the tree. **Default = None**. If None, then nodes are expanded until all leaves are pure or until all leaves contain less than min_samples_split samples.

min_samples_leaf – The minimum number of samples required to be at a leaf node. **Default = 1** Both left and right child nudes must contain min_samples_leaf training samples after splitting.

min_samples_split – The minimum number of samples required to split an internal node. **Default = 2**

Setting these three parameters to values different from the default settings was used for pruning the resulting decision tree, described further. However, for the initial run of the script, the parameters of the 'DecisionTreeRegressor' function were set to pseudo default settings to model a complex decision tree as a benchmark. The used pseudo default settings were as follows:

- **max_depth** – pseudo default value **= 6**
- **min_samples_leaf** – pseudo default value = default value **= 1**
- **min_samples_split** – pseudo default value = default value **= 2**

The reason for setting the **pseudo default value of max_depth to 6** differently from the default value is that the decision trees of a higher depth were growing into a size so large, that the exported pictures of these trees were no longer readable at a given resolution of the pictures, thus impossible to interpret. Moreover, the practical experience has shown, the pruned decision trees rarely had the depth of more than 4.

The Figure 20 shows the difference between un **unpruned tree** on the left, which is a result of a decision tree model with pseudo-default setting parameters **versus a pruned tree** on the right, from the same model after setting the limits on parameters for effective pruning. The examples of decision tree diagrams below are coming from the type A model of dependency of the transport mode Bike calculated from the nonfiltered school commute data, further abbreviated as A-School-Nonfiltered.

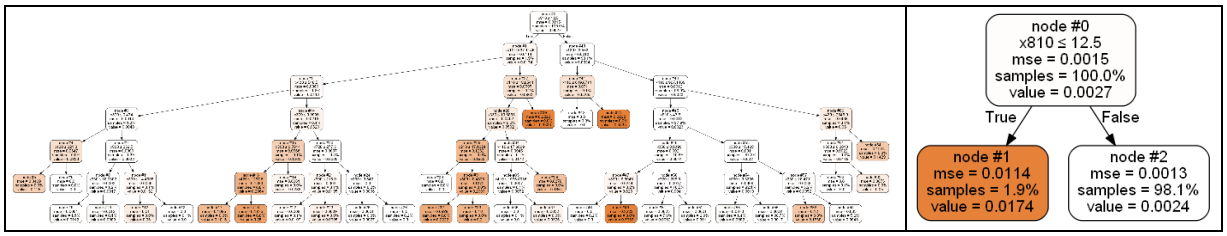


Figure 20: Example of graphs of unpruned tree on the left and pruned tree on the right, both taken from the A-School-Filtered decision tree model of the transport mode Bike

Description of the information in the decision tree diagram:

- Each node is numbered. The root node has **number** #0.
- The decision nodes include the **decision rule** – an inequality involving one of the explanatory variables and its respective threshold, which was calculated by the 'DecisionTreeRegressor' function to achieve the largest possible information gain at each decision node. For example: x810 ≤ 12.5 as in the pruned decision tree diagram in the Figure 20. Dependency intervals are then derived from the decision rule, for the case of pruned decision tree, value = 0.0174 for $x810 \in \langle 0; 12.5 \rangle$ and value = 0.0024 for $x810 \in (12.5; \infty)$.

- Each node also contains the information about:
    - **MSE** (mean square error) of the data in the node,
    - **number of samples** (data rows, in this case the O-D pairs) in the node,
    - average **value** of all Y data (the modal share) among the samples in the node.
- The nodes of the diagram are highlighted by a **colour scale** based on the value of the node (higher the value, the darker the colouring).

On one hand, **pruning the decision tree** decreases its complexity, which has positive impact on its interpretability as well as its generalisation, which should increase its applicability on various data set with similar data. On the other hand, pruning the tree usually has a negative impact on the root mean square error of the model.

The **root mean square error** (RMSE) is calculated simply as a square root of the mean square error (MSE) described in the theory (chapter 2.2.1). Square rooting the MSE is important to get characteristic of the model prediction error in scale comparable with the standard deviation (SD) of the input data. When creating or tuning the model, the general goal is to achieve:

1) RMSE of the training data model to be smaller than the SD of the training input data
2) RMSE of the training data model to be as small as possible
3) Complexity of the decision tree to be as small as possible

As already indicated, the above points 2) and 3) are commonly going against each other, thus optimal setting of model parameters is sometimes a result of an **effort to find balance** between those two. Example of SD of the training input data, RMSE of the training data model, depth of the tree and number of its leaves before and after pruning can be seen in Table 18 below. These outcomes are for the same model as the decision tree diagrams shown in Figure 20. In the example below, the RMSE after pruning is not as small as before, however, it is still smaller than the SD and results in significant reduction of the complexness of the decision tree.

Table 18: Example of standard deviation (SD) of the training input data, root mean square error (RMSE) of the training data model, depth of the tree and number of its leaves before and after pruning, taken from the A-School-Filtered model of the mode Bike

A-Nonfiltered-School		**Pseudo-default**			**Pruned**		
Mode	**SD**	**RMSE training**	**Depth**	**Leaves**	**RMSE training**	**Depth**	**Leaves**
Bike	0.04130	0.03730	6	33	0.03859	1	2

Another result of the tree pruning is decreasing the number of explanatory **variables important** within the model (used for explanation of the explained variable). This should be also obvious from the Figure 20, where the unpruned tree is split multiple times based on various explanatory variables, while the

pruned tree has split only once based on the variable x810 (Mountainousness of municipality). The other side of the reduction of number of important explanatory variables playing role in explaining the modal share of a transport mode is the increasing importance of the remaining explanatory variables. An example of importance of explanatory variables before and after the pruning can be seen in Table 19 below, again for the same model as in the previous examples.

Table 19: Example of importance of the explanatory variables (Table 9 can be used for their identification) before and after the pruning, taken from the A-School-Filtered model of the transport mode Bike

Default	**x110**	**x130**	**x210**	**x330**	**x360**	**x370**	**x410**	**x420**	**x430**	**x500**	**x720**	**x800**	**x810**	**Pruned**	**x810**
Bike	0.11	0.03	0.02	0.31	0.08	0.11	0.21	0.01	0.06	0.00	0.01	0.00	0.05	Bike	1.00

The **pruning** is **practically** done in several steps. Based on the goals 1) through 3) stipulated above, the first step is to find such parameter settings, which lead to local or absolute minimum of RMSE of the training data. Second step is to examine the impact of such settings to the complexness of a calculated decision tree. Third step is usually finding optimal settings, which balance the need for minimising the RMSE and pruning of the decision tree, while keeping in mind the goal to keep the RMSE of the training data to be smaller than the SD of the training input data.

The search for parameter settings, which lead to local or absolute minimum of RMSE of the training data starts by **plotting the course of RMSE** of the training data during the validation based on series of max_depth, min_samples_leaf and min_samples_split parameters settings. The course of RMSE in the case of max_depth has usually overall increasing tendency, example can be seen in Figure 21. The course of RMSE in the case of min_samples_leaf and min_samples_split is usually having the same tendency, either increasing or decreasing. Example of such can be seen in Figure 22.

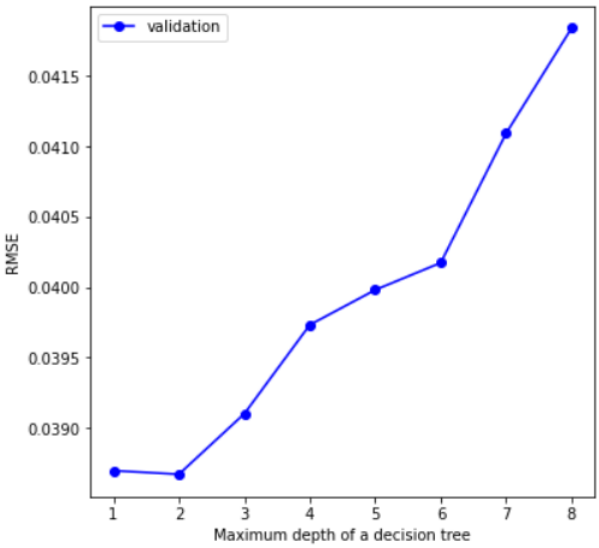


Figure 21: Course of RMSE of the training data during the validation based on a series of max_depth parameters settings, taken from the A-School-Filtered decision tree model of the transport mode Bike

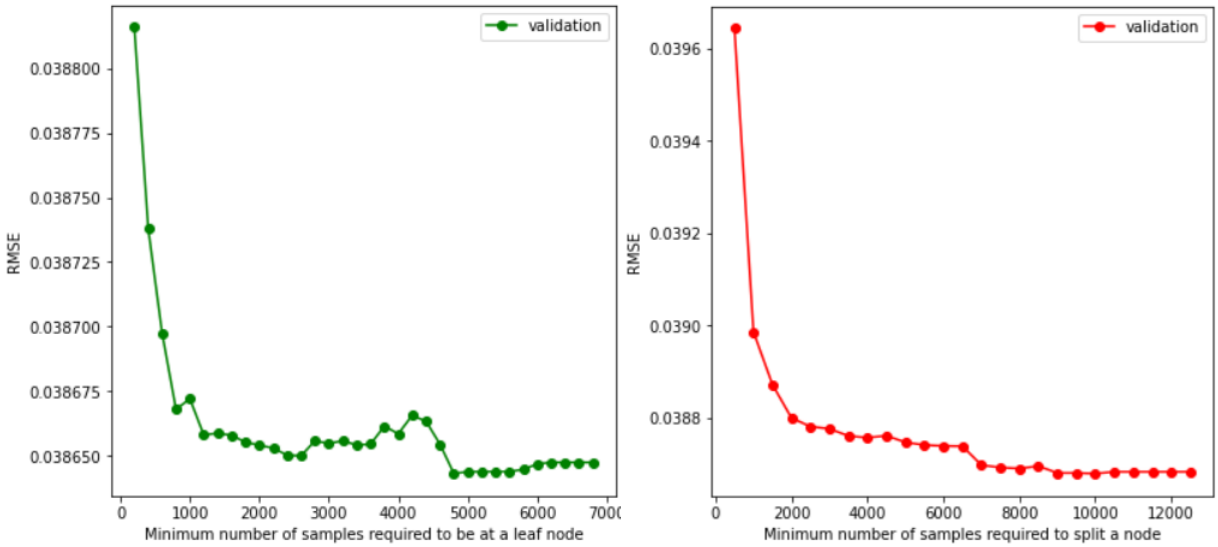


Figure 22: Course of RMSE of the training data during the validation based on series of min_samples_leaf (on the left) and min_samples_split (on the right) parameters settings, taken from the A-School-Filtered decision tree model of the transport mode Bike

The local and absolute minimums of max_depth, min_samples_leaf and min_samples_split read from the respective graphs are then recorded as an input to the '**GridSearchCV**' function, which re-calculates the decision tree model with all mutual combinations of the recorded parameter settings and recommends the best combination to be used for the calculation of the pruned decision tree. The combination can be further adjusted if needed to find optimal settings, which balances the need for minimising the RMSE and pruning of the decision tree, as explained above.

Once the parameter settings are optimised, final re-calculation of the model is run. Full set of outputs from the final re-calculation is exported and saved. Among others, **RMSE of the test data** model is also recorded in the end. Once RMSE of the test data model is recorded and put next to the RMSE of the training data model (both coming from a model with the same parameter settings used for pruning the tree) an additional verification of the difference between these two RMSE values can be made. If this final check proves the difference between the RMSE of training and test data models has decreased in the case of models using the optimal pruning parameters in comparison to models using the pseudo-default parameters setting, one can assume the pruning has contributed to a higher generalisation of the model and thus a higher reliability of its outcomes.

All **values relevant to the pruning process** of decision tree recorded during the modelling of each transport mode can be seen in Table 20 – the 'pruning table'. This set of values is an overview of how the performance of the decision tree model has changes from using the pseudo-default parameter settings to using the optimal pruning parameter settings.

Table 20: Pruning table – values relevant to pruning of decision tree recorded during modelling of each transport mode, Diff = difference between the root mean square error (RMSE) of training and test data models

A-Nonfiltered-School		Pseudo-default					Pruned				
		RMSE			Tree		RMSE			Tree	
Mode	SD	Training	Test	Diff	Depth	Leaves	Training	Test	Diff	Depth	Leaves
Bike	0.04130	0.03730	0.05131	27%	6	33	0.03859	0.05052	24%	1	2

The decision tree models can be remodelled using only the explanatory variables, which have any importance for the models or set of models. **The remodelling** with the reduced set of explanatory variables should be less computation time consuming and might result in change of values relevant to the pruning process as well as the importance of explanatory variables.

4.3.2 Linear regression modelling

The dependency of the modal split (the explained variable) on the explanatory variables was researched using linear regression models, theoretically described in chapter 2.2.2. The linear regression model provides a **Pearson correlation coefficient (r)**, which is indicating the strength of the linear correlation between the modal share and the explanatory variable. The Pearson correlation coefficient was calculated using the function 'correl' available in the MS Excel software.

The correlation coefficient was calculated based on values of the transport mode share in all O-D pairs in the respective data set and values of the explanatory variable in the O-D pairs. Thus, the strength of the linear correlation given by the coefficient is always a dependency between some single transport mode share and some **single explanatory variable** and it is not influenced by the dependency of the modal share on the other explanatory variables. Thus the correlation cannot be affected by the mutual dependency of explanatory variables.

The correlation coefficients were calculated for all combinations (all pairs) of transport mode shares and explanatory variables. The coefficients of one transport mode with all explanatory variables form a **vector of correlation coefficients**. Such a vector then describes the correlation of one transport mode share with all explanatory variables. One vector of correlation coefficients was filled for every transport mode. Multiple vectors for multiple transport modes can be then presented as a table. Such table was created for each filtered data set reflecting the model type (A, B or C) and commute purpose (work and school) and some nonfiltered data sets for comparison.

The correlation coefficient can take values from -1 to 1. The distance of the value from 0 is describing the strength of the correlation, while the negative/positive signs are providing the information about the negative/positive association. When only the information about the strength of the correlation is required, the **coefficient of determination (R^2)** can be used. The values of coefficient of determination

can be obtained by simple squaring all correlation coefficients in the vectors or tables of correlation coefficients. The coefficient of determination then takes values from 0 to 1.

The key **advantages of the coefficient of determination** are:

- its easier interpretation, when comparing different strengths of dependencies;
- its non-negative values allowing averaging;
- its range is from 0 to 1, which is a common range of ratio-based variables.

The coefficient of determination allows for averaging the dependency strength in the vectors or tables of coefficient of determination. It is then possible to **compare the performance of the various models** (type A, B and C) between each other and also the performance of models of the same type, which are using different subsets of explanatory variables.

On the other hand, **neither of the two coefficients can provide** a complex description of the modal share dependency on the explanatory variables, which would include the dependency intervals and decisive thresholds of the explanatory variables. To obtain those, the decision tree models were used.

As the coefficient of determination describes the strength of dependency of the modal share on the explanatory variable, it is in some sense comparable with the importance of the explanatory variable, which is describing the contribution of each explanatory variable to information gain of the decision tree model of each transport mode. The calculation of the coefficient of determination and of the importance of the explanatory variable follows different principles, as can be seen in chapters 2.2.1 and 2.2.2. However, they are both able **to sort the explanatory variables** according to their level of contribution to explanation of the modal share within the models. Thus, it is then possible, for example, to compare the set of most contributing variables based on both coefficient of determination and importance of the explanatory variable. To make such comparison meaningful, only linear regression and decision tree models using the same number of explanatory variables need to be compared. Moreover, each coefficient of determination should also be normalised to values between 0 and 1 by dividing by the sum of coefficients of determination of all used explanatory variables. The importance of the explanatory variable is already normalised on a similar principle, as can be seen in formula (9).

[5] Findings

This chapter will be devoted to a presentation, analysis and discussion of findings from the research stages with some findings. The stages are arranged in the order they have taken in the workflow.

5.1 Reduction of origin-destination pairs

The findings in this chapter are related to the change of modal split, travel time and average number of journeys within O-D pairs in the data sets **before and after the reduction**. During the reduction, O-D pairs from or to national and regional capitals as well as the O-D pairs between the municipalities with extended powers (MEP) were removed, leaving only the O-D pairs related to the local commute. The overview of the changes induced by removing the large towns and cities will indicate the differences between the local commute and commute to large towns and cities.

5.1.1 Presentation of findings

It is necessary to **define the presented modal split** first. It is the modal split of daily commute to work or school between known origins and destination, using known transport mode with known travel time. It is calculated from the total number of journeys in the data set, which is the same as the total number of commuters in the sample. No journeys taken by insignificant modes were removed from the data set, but they were aggregated into a residual transport mode 'the Rest'. The modal share of each transport mode is then expressed as a percentage of total number of journeys.

The comparison starts with the difference between **modal split of work commute**. The Table 21 is showing the modal split in data set, which includes both the local commute and the commute to large towns and cities. This data set was created from the original data set obtained from CSO by removing only a negligible number of incomplete records, see chapter 4.1.2. Whereas the Table 22 is presenting the modal split in data set, which includes only O-D pairs of local commute, see chapter 4.1.6.

Table 21: Modal split of work commute in the data set including O-D pairs from all over the nation

WORK nationwide	Bus	Train	MCT	Driver	CarPass	PTplus	Bike	PTcom	Walk	Rest	TOTAL
Number of journeys	127 860	37 252	36 626	466 435	74 998	76 364	18 002	52 786	9 949	26 199	926 471
Modal split	13.8%	4.0%	4.0%	50.3%	8.1%	8.2%	1.9%	5.7%	1.1%	2.8%	100.0%

Table 22: Modal split of work commute in the data set including only O-D pairs of the local commute

WORK local	Bus	Train	MCT	Driver	CarPass	PTplus	Bike	PTcom	Walk	Rest	TOTAL
Number of journeys	88 019	14 847	8 980	279 425	48 842	39 100	15 195	9 154	6 854	17 997	528 413
Modal split	16.7%	2.8%	1.7%	52.9%	9.2%	7.4%	2.9%	1.7%	1.3%	3.4%	100.0%

A comparison of difference between **modal split of school commute** follows. The Table 23 shows the modal split in the nationwide data set, while the Table 24 only in the local commute data set.

Table 23: Modal split of school commute in the data set including O-D pairs from all over the nation

SCHOOL nationwide	Bus	Train	MCT	Driver	CarPass	PTplus	Bike	PTcom	Walk	Rest	TOTAL
Number of journeys	122 662	25 424	28 427	9 481	27 104	24 437	910	50 604	5 773	828	295 650
Modal split	41.5%	8.6%	9.6%	3.2%	9.2%	8.3%	0.3%	17.1%	2.0%	0.3%	100.0%

Table 24: Modal split of school commute in the data set including only O-D pairs of the local commute

SCHOOL local	Bus	Train	MCT	Driver	CarPass	PTplus	Bike	PTcom	Walk	Rest	TOTAL
Number of journeys	94 496	11 893	7 356	3 915	18 093	12 709	661	11 865	2 950	415	164 353
Modal split	57.5%	7.2%	4.5%	2.4%	11.0%	7.7%	0.4%	7.2%	1.8%	0.3%	100.0%

Another travel behaviour phenomenon to be compared between the nationwide and the local data set is the **travel time**. It is important to note that the travel time went through two-stage averaging process. At first, the multiple travel time records relevant to one transport mode used within one O-D pair (coming from multiple commuters, i.e. census forms) were averaged to obtain a single average travel time by the one transport mode within each O-D pair, as described in chapter 4.1.4. In the second stage, the average travel times by the mode from all O-D pairs were averaged again to obtain the average travel time by the mode within the data set, which is presented in the tables below.

Work commute average travel times will be compared first. The Table 25 shows the average travel times in the nationwide data set, while the Table 26 only in the local commute data set.

Table 25: Average travel times by transport modes used for work commute on O-D pairs from all over the nation

WORK nationwide	Bus	Train	MCT	Driver	CarPass	PTplus	Bike	PTcom	Walk	Rest
Average travel time [min]	36.3	50.2	33.7	28.8	26.1	38.4	26.5	60.8	20.3	26.1

Table 26: Average travel times by transport modes used for work commute on O-D pairs of the local commute

WORK local	Bus	Train	MCT	Driver	CarPass	PTplus	Bike	PTcom	Walk	Rest
Average travel time [min]	32.5	42.0	31.2	25.7	22.5	31.7	24.4	52.4	21.4	22.8

A comparison of difference between **school commute average travel times** follows. The Table 27 shows the average travel times in the nationwide data set, the Table 28 of the local commute.

Table 27: Average travel times by modes used for school commute on O-D pairs from all over the nation

SCHOOL nationwide	Bus	Train	MCT	Driver	CarPass	PTplus	Bike	PTcom	Walk	Rest
Average travel time [min]	36.6	53.3	28.1	28.9	19.6	38.7	27.7	60.8	15.2	28.2

Table 28: Average travel times by transport modes used for school commute on O-D of the local commute

SCHOOL local	Bus	Train	MCT	Driver	CarPass	PTplus	Bike	PTcom	Walk	Rest
Average travel time [min]	32.6	45.1	24.8	23.1	16.0	29.9	22.0	52.5	15.5	21.5

5.1.2 Analysis of findings

The **change of modal split** induced by the reduction of O-D pairs is presented in Table 29 (work commute) and Table 30 (school commute). The tables put the previously presented modal split figures into one place for easier recognition of the change between the nationwide and local data sets. Transport modes that have seen an increase in their modal share after the reduction are marked with a plus sign, decreasing modes with a minus sign.

Table 29: Change of modal split of work commute between data sets of nationwide and local O-D pairs

WORK nationwide → local	**Bus**	**Train**	**MCT**	**Driver**	**CarPass**	**PTplus**	**Bike**	**PTcom**	**Walk**
Modal split nationwide	13.8%	4.0%	4.0%	50.3%	8.1%	8.2%	1.9%	5.7%	1.1%
Modal split local	16.7%	2.8%	1.7%	52.9%	9.2%	7.4%	2.9%	1.7%	1.3%
Change of modal split	+	-	-	+	+	-	+	-	+

Table 30: Change of modal split of school commute between data sets of nationwide and local O-D pairs

SCHOOL nationwide → local	**Bus**	**Train**	**MCT**	**Driver**	**CarPass**	**PTplus**	**Bike**	**PTcom**	**Walk**
Modal split nationwide	41.5%	8.6%	9.6%	3.2%	9.2%	8.3%	0.3%	17.1%	2.0%
Modal split local	57.5%	7.2%	4.5%	2.4%	11.0%	7.7%	0.4%	7.2%	1.8%
Change of modal split	+	-	-	-	+	-	+	-	-

Except the transport modes Driver and Walk, **basically all modes have shown the same trend** of the change due to reduction in both work and school commute. However, the transport mode Walk has shown a minimum change in both of the cases, thus it can be considered rather invariant. This invariance is likely caused by an equal availability of this mode in all areas of Czechia.

The transport modes **Train, MCT, PTplus** and **PTcom** have shown a **decrease** in their modal shares, after the O-D pairs to and from and from large towns and cities were removed from the data sets. That implies, those modes are of more importance when it comes to commute to and from the large towns and cities. This is in line with increasing passenger volumes in suburban trains, which are repeatedly reported in past years, as well with the fact the mass city transport systems are set up mainly in the large towns and cities. That would also imply that the combined modes PTplus and PTcom involve more trips made by Train and MCT (which are the modes with the decreased share) than by Bus (which is the mode with the increased share).

On the contrary, the modes **Bus, Driver, CarPass** and **Bike** has shown an increase of their modal shares, after the O-D pairs to and from and from large towns and cities were removed. That implies higher importance of these modes for the local commute. This would be in line with their general availability, which is giving them the comparative advantage in areas uncovered by the Train or MCT services, which are typically the rural or remote areas of the country.

In terms of the **travel time changes**, similarly to the modal split, the transport mode Walk is the distinct one. All transport modes have shown a decrease in the travel time but Walk. The decrease in the travel times of nearly all transport modes after the reduction of O-D pairs to and from the large towns and cities implies that the commute to and from large towns and cities takes more time. Based on the strong correlation of the travel time with the travel distance, the longer travel times also indicate larger catchment areas for commuting to large towns and cities, which is in line with the studies on their size, for instance the reviewed Mulíček et al. (2013) and Čekal (2006).

Another comparable metric of the data sets before and after the reduction is the **average number of journeys within one O-D pair**. In the case of work commute, the average number of journeys within one O-D pair has dropped from 7.98 to 5.39 after removing the O-D pairs to and from the large towns and cities. In other words, the O-D pairs of the local commute are less busy, than the O-D pairs involving the large towns and cities, which makes perfect sense. It is confirmed by the decrease in the average number of journeys within one O-D pair of school commute from 7.35 to 5.85.

It was necessary to remove the O-D pairs to and from the large towns and cities from the data sets to obtain the data set for local commute, which is the subject of this research. Moreover, the comparison of the **nationwide and local commute** has shown these two **differ significantly** and to do a research of one or another is a different discipline. This confirms the correctness of the decision to focus only one of them, i.e. the local commute. Therefore, the explanatory variables were tailored to the local commute and their dependency on the nationwide commute cannot and will not be researched.

5.2 Filtration of origin-destination pairs

This stage included removing selected O-D pairs from already reduced data set of O-D pairs related to the local commute. The **reasoning for filtration** is to reduce the distortion of the modal split in the data sets caused by the O-D pairs with small number of travellers, typically only one traveller. The 100 % share of the transport mode used by the single traveller within the O-D pair is an extremely high value, which is on the other hand justified by extremely small number of travellers, namely only one. Thus, it is desirable to increase the ratio of modal shares supported by larger numbers of travellers within the data sets. This will increase the reliability of the overall modal split figures.

The filtration was done using the $\boldsymbol{V_{ij}}$ **filtration method**, which is removing the O-D pairs of lesser importance from the data sets and which was assessed the best preforming in chapter 4.1.7. The importance of the O-D pairs was assessed based on the share the O-D pairs have on the total in- and out-commuting journeys from origin and to destination municipalities. The O-D pairs used by less commuters are then having higher probability to become classified unimportant, thus removable.

The findings in this chapter are related to the change of modal split, travel time and average number of journeys within O-D pairs in the data sets **before and after the filtration**. The overview of the changes, induced by removing the O-D pairs with smaller share (importance) on the total in- and out-commuting flows in the municipalities of origin and destination, will indicate the differences between the nonfiltered and filtered data sets of O-D pairs.

In this chapter, the designation **'nonfiltered' data set is used for** the otherwise unaltered **'local' data sets** from the previous chapter. They are compared with 'filtered' data sets (i.e. the outcome of filtration of the local data sets). The definitions of the presented modal split, travel time and average number of journeys within one O-D pair remain the same as in the previous chapter.

5.2.1 Presentation of findings

The comparison starts with the difference between **modal splits of work commute**. Table 31 is showing the modal split in the nonfiltered data set of O-D pairs, whereas Table 32 is presenting the modal split in the filtered data set of O-D pairs.

Table 31: Modal split of work commute in the nonfiltered data set of O-D pairs

WORK nonfiltered	Bus	Train	MCT	Driver	CarPass	PTplus	Bike	PTcom	Walk	Rest	TOTAL
Number of journeys	88 019	14 847	8 980	279 425	48 842	39 100	15 195	9 154	6 854	17 997	528 413
Modal split	16.7%	2.8%	1.7%	52.9%	9.2%	7.4%	2.9%	1.7%	1.3%	3.4%	100.0%

Table 32: Modal split of work commute in the filtered data set of O-D pairs

WORK filtered	Bus	Train	MCT	Driver	CarPass	PTplus	Bike	PTcom	Walk	Rest	TOTAL
Number of journeys	78 319	10 952	8 144	207 826	39 918	33 076	13 757	6 547	5 773	14 375	418 687
Modal split	18.7%	2.6%	1.9%	49.6%	9.5%	7.9%	3.3%	1.6%	1.4%	3.4%	100.0%

A comparison of difference between **modal split of school commute** follows. Table 33 shows the modal split in the nonfiltered data set, while Table 34 in the filtered data set.

Table 33: Modal split of school commute in the nonfiltered data set of O-D pairs

SCHOOL nonfiltered	Bus	Train	MCT	Driver	CarPass	PTplus	Bike	PTcom	Walk	Rest	TOTAL
Number of journeys	94 496	11 893	7 356	3 915	18 093	12 709	661	11 865	2 950	415	164 353
Modal split	57.5%	7.2%	4.5%	2.4%	11.0%	7.7%	0.4%	7.2%	1.8%	0.3%	100.0%

Table 34: Modal split of school commute in the filtered data set of O-D pairs

SCHOOL filtered	Bus	Train	MCT	Driver	CarPass	PTplus	Bike	PTcom	Walk	Rest	TOTAL
Number of journeys	84 946	8 491	6 797	2 729	16 274	10 769	618	7 798	2 226	330	140 978
Modal split	60.3%	6.0%	4.8%	1.9%	11.5%	7.6%	0.4%	5.5%	1.6%	0.2%	100.0%

Next comparison will be devoted to travel times. **Work commute average travel times** will be compared first. Table 35 shows the average travel times in the nonfiltered data set, while Table 36 in the filtered data set.

Table 35: Average travel times by transport modes used for work commute in the nonfiltered data set

WORK nonfiltered	Bus	Train	MCT	Driver	CarPass	PTplus	Bike	PTcom	Walk	Rest
Average travel time [min]	32.5	42.0	31.2	25.7	22.5	31.7	24.4	52.4	21.4	22.8

Table 36: Average travel times by transport modes used for work commute in the filtered data set

WORK filtered	Bus	Train	MCT	Driver	CarPass	PTplus	Bike	PTcom	Walk	Rest
Average travel time [min]	27.9	33.4	26.0	18.5	18.5	25.6	23.0	42.1	22.7	19.1

A comparison of difference between **school commute average travel times** follows. Table 37 shows the average travel times in the nonfiltered data set, Table 38 in the filtered.

Table 37: Average travel times by transport modes used for school commute in the nonfiltered data set

SCHOOL nonfiltered	Bus	Train	MCT	Driver	CarPass	PTplus	Bike	PTcom	Walk	Rest
Average travel time [min]	32.6	45.1	24.8	23.1	16.0	29.9	22.0	52.5	15.5	21.5

Table 38: Average travel times by transport modes used for school commute in the filtered data set

SCHOOL filtered	Bus	Train	MCT	Driver	CarPass	PTplus	Bike	PTcom	Walk	Rest
Average travel time [min]	26.6	35.0	22.2	18.7	14.4	24.1	20.9	41.1	19.2	18.4

5.2.2 Analysis of findings

The **change of modal split** induced by the filtration of O-D pairs is presented in Table 39 (work commute) and Table 40 (school commute). Transport modes that have seen an increase in their modal share after the filtration are marked with a plus sign, decreasing modes with a minus sign.

Table 39: Change of modal split of work commute between nonfiltered and filtered data sets

WORK nonfiltered → filtered	Bus	Train	MCT	Driver	CarPass	PTplus	Bike	PTcom	Walk
Modal split nonfiltered	16.7%	2.8%	1.7%	52.9%	9.2%	7.4%	2.9%	1.7%	1.3%
Modal split filtered	18.7%	2.6%	1.9%	49.6%	9.5%	7.9%	3.3%	1.6%	1.4%
Change of modal split	+	-	+	-	+	+	+	-	+

Table 40: Change of modal split of school commute between nonfiltered and filtered data sets

SCHOOL nonfiltered → filtered	Bus	Train	MCT	Driver	CarPass	PTplus	Bike	PTcom	Walk
Modal split nonfiltered	57.5%	7.2%	4.5%	2.4%	11.0%	7.7%	0.4%	7.2%	1.8%
Modal split filtered	60.3%	6.0%	4.8%	1.9%	11.5%	7.6%	0.4%	5.5%	1.6%
Change of modal split	+	-	+	-	+	-	+	-	-

The first thing to note about the **change of modal split** after the filtration is that it **is much smaller** than the change, which has occurred after the reduction, analysed in chapter 5.1.2.

Due to the relatively small change of modal split, **it is difficult to clearly identify trends**, which would be common to both work and school commute. In fact, no transport mode is showing large change in its modal share in both work and school commute. Also, because the modal shares of more transport modes are rather stagnating after the filtration. Therefore, only changes to the modal share, which are of a larger extent and with the same sign in both work and school commute will be analysed.

The filtration is based on removal of the less important O-D pairs. An O-D pair has a lower importance if it is used by a small share of commuters, compared to the total number of commuters, either leaving the municipality of origin or entering the municipality of destination. A commuting tie of the two municipalities, which are connected by the O-D pair of lower importance is less important as well. The reason for **lower importance of the commuting tie** of the two municipalities can be given either by their greater mutual distance or by a lower mutual attractivity for their inhabitants. Both of these factors negatively affect the extent of public transport services offered between those municipalities, which could result in increasing use of some private transport.

Based on the modal change comparison, the transport mode **Driver** is the one, which is having higher share in the data sets including unimportant O-D pairs and whose share decreases after filtering those out of the data sets. This is in line with the expected higher share of the private transport modes in the nonfiltered data sets including more O-D pairs of lower importance.

Also the transport mode **PTcom** decreases after filtering out the unimportant O-D pairs. The average travel time by the mode PTcom is the longest in all assessed data sets. Based on a high correlation of the travel time and the travel distance, it can be assumed the mode PTcom is used for longer travel distances, which are the most affected by the filtering based on the lower importance. Of course, the travel time – travel distance relation of the mode PTcom could be biased by the transfer waiting time, which is included in the total travel time of the combined mode PTcom.

The public transport modes are generally expected to increase in the filtered data set of more important (more used) O-D pairs. However, the mode **Train** is decreasing, which was rather unexpected. The reason for that could be in a low flexibility of the mode Train. Most of the railway network in Czechia was built before the World War I, thus over a hundred years ago. Since that time, the importance of some local commuting ties might have changed, however, unlike the other public transport modes, the Train services cannot be easily rerouted due to high cost of rebuilding its infrastructure.

Modal shares of public transport modes **Bus** and **MCT** has increased after the filtration as expected.

The private transport modes are generally expected to decrease after filtering out the less important (less used) O-D pair, where public transport is not available or less competitive. However, the mode **Bike** is increasing, which was rather unexpected. The explanation could be related to the factors, which are influencing the use of mode Bike. The research of dependencies of modal shares of transport modes on explanatory variables has shown that the modal share of mode Bike is increasing with the increasing population size of the municipality of destination, see chapter 5.7. The O-D pair to municipalities with higher population are likely to be assessed as important, because such municipalities are more likely to be an attractive commute destination.

In terms of the **travel time changes**, the recorded changes are similar to ones recorded after the reduction. All transport modes has shown a decrease in the travel time but Walk. The decrease in the travel times of nearly all transport modes after the filtration of O-D pairs of lower importance implies that the commute within the less important O-D pairs takes more time. That is in line with the assumption that the mutual distance of municipalities has a negative effect on the importance of the commuting tie between them and thus also importance of the relevant O-D pair.

Another comparable metric of the data sets before and after the reduction is the **average number of journeys within one O-D pair**. In the case of work commute, the average number of journeys within one O-D pair has increased from 5.39 to 11.193 after removing the less important (less used) O-D pairs, which makes perfect sense. It is confirmed by the increase of average number of journeys within one O-D pair of school commute from 5.85 to 9.57. In the end, the resulting average numbers of journeys within one O-D pair in both data set are higher than in the original nationwide data set before reduction (7.98 work commute and 7.35 school commute). The increased averages allow to focus in the research on the most important O-D pairs.

Another benefit of filtering out the less important O-D pairs is the decreased spread of the modal share data. This can be demonstrated by the **box and whisker plots**, comparing the structure of Bus modal share data related to school commute before and after the filtration, presented in Figure 23. The box and whisker plots for all transport modes concerning both work and school commute are listed in the annex [D]. Few tips to interpret the box and whisker plots follow, based on the explanation provided by Excel Easy (BoxAndWhisherPlot):

- the middle line in the box represents the median (of all modal share data),
- the 'X' in the box represents the mean of the data,
- the bottom line of the box represents the 1st quartile (Q_1),
- the top line of the box represents the 3rd quartile (Q_3),
- the distance between the box bottom and top line is referred as an interquartile range (IQR),

- the bottom line of the whisker represents the 'minimum' (min value, which is not an outlier),
- the top line of the whisker represents the 'maximum' (max value, which is not an outlier),
- any dots below the bottom line of the whisker are outliers if they are below Q_1 – 1.5 * IQR,
- any dots above the top line of the whisker are outliers if they are above Q_3 +1.5 * IQR.

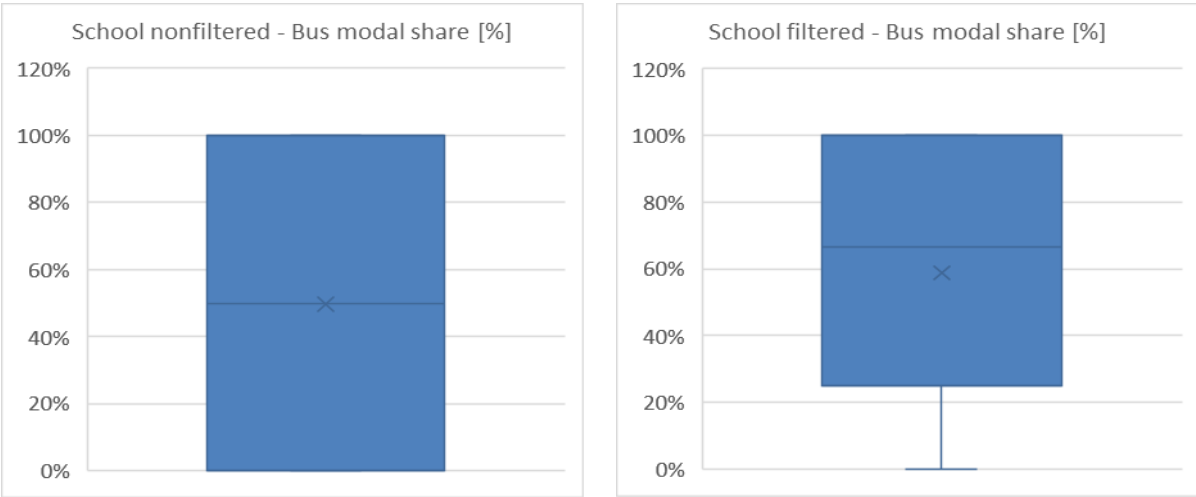


Figure 23: Comparison of the structure of data on modal share of Bus in school commute before filtration (on the left) and after filtration (on the right)

The ultimate purpose of the filtration is to filter out (to remove) the O-D pairs used by a little number of commuters, especially the O-D pairs with only one traveller, as described in chapter 4.1.7. Such O-D pairs can cause a distortion of the modal split figures and ultimately even of the **dependencies of the modal share on the explanatory variables**. Thus, it is necessary to verify the overall change in the strength of dependencies of modal shares. This was done using the linear regression model described in chapter 2.2.2 and resulting tables of coefficient of determination, described in chapter 0.

The coefficients of determination were calculated for all combination (all pairs) of transport mode shares and explanatory variables. The coefficients of one transport mode with all explanatory variables form a vector of coefficients of determination. The vectors made for all transport modes then form a **table of coefficient of determination** for the particular data set. In total 4 tables of coefficients of determination were produced for the type A model, as the average dependency strength in nonfiltered and filtered data sets was compared for both work and school commute. The full tables of coefficients of determination are presented in annex [E]. The tables below present the result of averaging of the coefficients of determination in the full tables per each transport mode as well as in the whole data sets.

Table 41 allows for comparison of the average coefficients of determination in the data sets of work commute before and after the filtration. The **average coefficients of determination of the whole data set** are describing the average strength of the dependency of the modal share on the explanatory

variables in each data set. The value of average coefficients of determination of the whole data set has increased from 0.00077 to 0.00120 after the filtration in the case of work commute data sets. Table 42 allows for comparison of the average coefficients of determination in the data sets of school commute before and after the filtration. The value has increased from 0.00270 to 0.00425 after the filtration in the case of school commute data sets. The tables use colour scale to emphasise the largest values of average coefficients of determination (green tones) and the lowest values (red tones).

Table 41: Average coefficients of determination (R^2) of each transport mode as well as of the whole data set, comparison of average R^2 in nonfilter and filtered data sets of work commute used for type A models

Work commute		
MODE	**AVERAGE R^2**	
	Nonfiltered	**Filtered**
Bus	0.00152	0.00153
Train	0.00097	0.00116
MCT	0.00120	0.00355
Driver	0.00122	0.00161
CarPass	0.00043	0.00055
PTplus	0.00018	0.00034
Bike	0.00050	0.00122
PTcom	0.00073	0.00065
Walk	0.00015	0.00017
DATA SET	**0.00077**	**0.00120**

Table 42: Average coefficients of determination (R^2) of each transport mode as well as of the whole data set, comparison of average R2 in nonfilter and filtered data sets of school commute used for type A models

School commute		
MODE	**AVERAGE R^2**	
	Nonfiltered	**Filtered**
Bus	0.00983	0.01510
Train	0.00339	0.00338
MCT	0.00431	0.01096
Driver	0.00026	0.00108
CarPass	0.00227	0.00304
PTplus	0.00041	0.00037
Bike	0.00011	0.00025
PTcom	0.00186	0.00161
Walk	0.00189	0.00249
DATA SET	**0.00270**	**0.00425**

The average coefficients of determination of both data sets have nearly doubled, which indicates a remarkable **increase of overall dependency strength** in both data sets after the filtration. In terms of the individual transport modes, most of them also show the increase of overall dependency strength. The exceptions are the transport modes Train and PTplus in the case of school commute and PTcom in the case of both school and work commute. However, the decrease of mode Train is negligible. A negative effect of the filtration on the dependency strength of mode PTcom might be connected to higher share of PTcom within O-D pairs, which require longer travel distances, and which are affected by the filtration.

Considering the increase of overall dependency strength is much higher than the changes to actual modal split figures presented before, the **filtration meets the expectations**. The negative effect on combined mode PTcom is an acceptable price for the use of filtration and increase of the overall dependency strength in all data sets.

5.3 Type A models of school commute

This stage included modelling dependencies of the school commute modal shares (explained variable) on the characteristics (explanatory variables) of the **municipalities of origin** (municipalities A), using the decision tree and the linear regression models. Any of these models can be referred as a type A model, because they all use the characteristics of municipalities A as the explanatory variables.

5.3.1 Presentation of findings

As stipulated in the methodology, the modelling started with the **type A decision tree models of school commute using the filtered data** set of O-D pairs (A-School-Filtered). In the first set of 9 models (**version 1** or simply v1), all 39 municipality-related explanatory variables were used. The overview of how the performance of the decision tree models has changed from using the pseudo-default parameter settings to using the optimal pruning parameter settings can be seen in Table 43, i.e. in the so called 'pruning table'. The pruning table also includes the percentage of decrease from the standard deviation (SD) to the root mean square error (RMSE) of training data model (SD→Tr). The greater is the decrease from SD to RMSE, the better. Positive values of SD→Tr indicate an increase of RMSE in comparison with SD, which is not desirable. A conditional formatting using three colour scale is applied to SD→Tr, and to the difference between the RMSE resulting from training and from test data models (Diff), to compare and highlight the performance of the decision tree models.

Table 43: Pruning table – type A decision tree models of school commute using the filtered set of O-D pairs, version 1, SD = standard deviation, RMSE = root mean square error, SD→Tr = decrease from the SD to the RMSE of the training data model, Diff = difference between the RMSE resulting from training and from test data models, D = depth of the decision tree, L = number of leaves of the decision tree

A-School-Filtered v1		Pseudo-default						Pruned					
		RMSE				Tree		RMSE				Tree	
Mode	SD	SD→Tr	Training	Test	Diff	D	L	SD→Tr	Training	Test	Diff	D	L
Bus	0.38152	-8.7%	0.34830	0.36010	3%	6	62	-6.3%	0.35764	0.35759	0%	5	16
Train	0.17844	-11.2%	0.15847	0.16112	2%	6	51	-7.2%	0.16554	0.15939	4%	4	11
MCT	0.10852	-12.9%	0.09455	0.10956	14%	6	49	-5.7%	0.10229	0.10822	5%	4	8
Driver	0.12287	-3.1%	0.11902	0.12459	4%	6	35	-0.1%	0.12271	0.12045	2%	3	5
CarPass	0.26991	-3.9%	0.25940	0.27336	5%	6	54	-1.5%	0.26578	0.27029	2%	4	10
PTplus	0.17018	-0.5%	0.16937	0.16708	1%	6	32	1.5%	0.17266	0.15878	8%	1	2
Bike	0.03926	-4.9%	0.03733	0.03531	5%	6	34	4.0%	0.04084	0.03156	23%	1	2
PTcom	0.17306	-4.8%	0.16472	0.18305	10%	6	39	-1.9%	0.16984	0.17654	4%	3	7
Walk	0.11938	-4.5%	0.11406	0.13292	14%	6	24	-2.1%	0.11690	0.12502	6%	2	4

As can be seen in Table 43, the pruned decision trees of transport **modes PTplus and Bike** are showing an increase in the RMSE values of pruned training data models in comparison with SD of the training data. For these two decision tree models, it was not possible to find any parameter settings, which

would effectively prune the decision tree and at the same time keep the RMSE of pruned training data models smaller than the SD of the training data. Thus, a decision was made to set the maximum tree depth of those two decision trees to 1, which also limits the number of used explanatory variables to 1. Because of the generally poor performance of the decision tree models, it is better to obtain a single erroneous importance of explanatory variable than a series of unreliable dependencies. After making all 9 A-School-Filtered decision tree models in version 1 (using all 39 explanatory variables), an overview of importance of explanatory variables was made. The overview has shown that 19 explanatory variables had no importance for any transport mode, i.e. did not contribute to information gain in any of the models. Those 19 variables are highlighted in red in the Table 46 and were removed from the set of explanatory variables used in the version 2 of A-School-Filtered decision tree models.

The pruning table of the **version 2** of the A-School-Filtered decision tree models can be seen in Table 44. The problem with the poor performance of PTplus and Bike models prevails, so it was solved in the same way by pruning the maximum tree depth to 1. Out of the 20 used explanatory variables, 3 of them had no importance for any transport mode, after making all 9 models. Those 3 variables are highlighted in brown in Table 46 and were removed from the set of explanatory variables used in the version 3 of A-School-Filtered decision tree models.

Table 44: Pruning table – type A decision tree models of school commute using the filtered set of O-D pairs, version 2, SD = standard deviation, RMSE = root mean square error, SD→Tr = decrease from the SD to the RMSE of the training data model, Diff = difference between the RMSE resulting from training and from test data models, D = depth of the decision tree, L = number of leaves of the decision tree

A-School-Filtered v2		Pseudo-default						Pruned					
		RMSE				Tree		RMSE				Tree	
Mode	SD	SD→Tr	Training	Test	Diff	D	L	SD→Tr	Training	Test	Diff	D	L
Bus	0.38152	-8.6%	0.34861	0.35801	3%	6	64	-6.4%	0.35708	0.35789	0%	6	19
Train	0.17844	-10.9%	0.15896	0.15928	0%	6	52	-7.0%	0.16601	0.15961	4%	3	8
MCT	0.10852	-13.2%	0.09422	0.10747	12%	6	53	-7.2%	0.10072	0.10682	6%	5	14
Driver	0.12287	-3.1%	0.11911	0.12409	4%	6	42	-0.1%	0.12278	0.12041	2%	3	5
CarPass	0.26991	-3.7%	0.25992	0.27237	5%	6	55	-1.7%	0.26533	0.26885	1%	4	10
PTplus	0.17018	-0.1%	0.17002	0.15761	7%	6	34	1.5%	0.17266	0.15878	8%	1	2
Bike	0.03926	-3.4%	0.03792	0.03368	11%	6	33	4.0%	0.04084	0.03156	23%	1	2
PTcom	0.17306	-4.8%	0.16475	0.18075	9%	6	42	-2.1%	0.16951	0.17678	4%	4	10
Walk	0.11938	-4.9%	0.11356	0.12916	12%	6	28	-2.0%	0.11704	0.12504	6%	2	3

The pruning table of the **version 3** of A-School-Filtered decision tree models can be seen in Table 45. The problem with the poor performance of PTplus and Bike models prevails, so it was solved in the same way by pruning the maximum tree depth to 1. All of the 17 used explanatory variables have some importance for at least one transport mode. Those 17 variables are the remaining variables suitable for application in linear regression models. They are stipulated in Table 46 without any colour.

Table 45: Pruning table – type A decision tree models of school commute using the filtered set of O-D pairs, version 3, SD = standard deviation, RMSE = root mean square error, SD→Tr = decrease from the SD to the RMSE of the training data model, Diff = difference between the RMSE resulting from training and from test data models, D = depth of the decision tree, L = number of leaves of the decision tree

A-School-Filtered v3		Pseudo-default						Pruned					
		RMSE				Tree		RMSE				Tree	
Mode	SD	SD→Tr	Training	Test	Diff	D	L	SD→Tr	Training	Test	Diff	D	L
Bus	0.38152	-8.6%	0.34856	0.35705	2%	6	64	-5.8%	0.35954	0.36091	0%	3	8
Train	0.17844	-10.8%	0.15913	0.15930	0%	6	53	-7.0%	0.16601	0.15961	4%	3	8
MCT	0.10852	-13.3%	0.09413	0.10727	12%	6	52	-7.2%	0.10072	0.10682	6%	5	14
Driver	0.12287	-3.1%	0.11912	0.12409	4%	6	42	-0.1%	0.12278	0.12041	2%	3	5
CarPass	0.26991	-3.7%	0.25980	0.27293	5%	6	55	-1.7%	0.26533	0.26885	1%	4	10
PTplus	0.17018	-0.1%	0.17005	0.15762	7%	6	34	1.5%	0.17268	0.15866	8%	1	2
Bike	0.03926	-3.4%	0.03792	0.03423	10%	6	33	4.0%	0.04084	0.03156	23%	1	2
PTcom	0.17306	-4.8%	0.16480	0.18152	9%	6	42	-2.0%	0.16954	0.17681	4%	4	10
Walk	0.11938	-4.7%	0.11383	0.13337	15%	6	27	-2.0%	0.11702	0.12499	6%	3	4

Table 46: Table of municipality-related explanatory variables, dropped variables highlighted in red and brown

Code	Group	Explanatory variable
x100	Age structure	Average age of inhabitants
x110	Age structure	Share of children aged 0-14
x120	Age structure	Share of inhabitants aged 15-64
x130	Age structure	Share of inhabitants with age 65+
x140	Age structure	Age index (x130/x110)
x200	Education	Share of employees with elementary education
x210	Education	Share of employees with apprenticeship certificate
x220	Education	Share of employees without high school diploma
x230	Education	Share of employees with high school diploma
x240	Education	Share of employees with completed extension study
x250	Education	Share of employees with tertiary technical education
x260	Education	Share of employees with university degree
x270	Education	Share of employees with high school diploma or higher
x280	Education	Average length of the education attendance of employees
x300	Econ. activity	Unemployment rate
x310	Econ. activity	Share of economically active inhabitants
x320	Econ. activity	Share of the retired
x330	Econ. activity	Share of pupils and students
x340	Econ. activity	Ratio of economically inactive to active inhabitants
x350	Econ. activity	Ratio of self-employed to economically active inhabitants
x360	Econ. activity	Ratio of occupied jobs to economically active inhabitants
x370	Econ. activity	Ratio of inhabitants employed in key sectors (T+/PS-)
x400	Tax revenue	Share of municipal inhabitant on national tax revenue
x410	Car ownership	Car ownership
x420	Population size	Number of inhabitants
x430	Population size	Total number of workers, pupils, students and apprentices
x500	Pop. density	Gross population density

Code	Group	Explanatory variable
x510	Pop. density	Net population density
x600	Trans. services	Municipality organizing Mass City Transport
x610	Trans. services	Distance to the closest train stop or station
x700	Real estate	Average purchase price of house
x710	Real estate	Average purchase price of apartment
x720	Real estate	Average purchase price of building land
x730	Real estate	Share of unoccupied dwellings
x800	Geomorphology	Mean altitude of municipality
x810	Geomorphology	Mountainousness of municipality
x820	Geomorphology	Mountainousness of municipality relative to its area
x905	Climate	Annual average precipitation
x915	Climate	Annual average temperature

The **strength of dependency** of the modal shares of transport modes on the set of 17 explanatory variables used in the version 3 of A-School-Filtered models is described by outputs of two models. The first is the importance of the explanatory variables from the decision tree model (DTR-IEV) and the second is the normalised coefficient of determination from the linear regression model (LR-R^2N). Both values are presented for each 'transport mode – explanatory variable' pair in the Table 47. **Not the absolute values of DTR-IEV and LR-R^2N shall be compared**, but the order of these from the largest to the smallest for both models of every transport mode.

Table 47: Importance of the explanatory variables (DTR-IEV) and normalised coefficients of determination (LR-R^2N) in the version 3 of A-School-Filtered models (using 17 explanatory variables). Table 46 can be used for identification of the variables. Decision tree models of modes PTplus and Bike have poor performance and were pruned to maximum depth of 1, see the Table 45. Top 1 (green), Top 2 (yellow) and Top 3 (salmon) values of dependency strength are highlighted in each row (transport mode model).

A-S-F v3	Model	x110	x130	x210	x320	x330	x360	x370	x410	x420	x430	x500	x610	x710	x720	x730	x800	x810
Bus	DTR-IEV	0	0	0.09	0	0	0	0.17	0	0	0	0.03	0.72	0	0	0	0	0
	LR-R^2N	0.01	0.02	0.16	0.04	0.01	0.01	0.12	0	0.06	0.06	0.10	0.21	0.03	0.10	0.04	0.02	0
Train	DTR-IEV	0	0	0.01	0	0	0.05	0	0.03	0.02	0	0	0.88	0	0	0	0	0
	LR-R^2N	0	0	0.04	0	0	0.01	0.03	0.10	0.01	0.01	0.04	0.62	0.02	0.02	0.08	0.03	0
MCT	DTR-IEV	0	0.09	0	0.01	0	0.08	0.36	0	0	0.05	0	0	0.29	0.02	0.09	0	0
	LR-R^2N	0.03	0.06	0.11	0.08	0	0.01	0.15	0	0.06	0.06	0.15	0.02	0.03	0.15	0.07	0.02	0
Driver	DTR-IEV	0	0	0	0	0.08	0	0	0.19	0.10	0.62	0	0	0	0	0	0	0
	LR-R^2N	0.02	0	0.08	0	0.02	0.04	0.02	0.04	0.24	0.24	0.13	0.01	0.02	0.10	0.02	0	0.01
CarPass	DTR-IEV	0	0	0.40	0	0.03	0	0	0.28	0.14	0	0	0	0	0.15	0	0	0
	LR-R^2N	0.03	0.02	0.24	0.04	0.06	0	0.18	0.1	0.02	0.02	0.02	0.06	0.04	0.14	0	0	0.01
PTplus	DTR-IEV	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
	LR-R^2N	0.02	0.01	0.06	0.01	0.01	0.03	0	0.22	0.06	0.07	0.07	0.03	0.07	0.07	0.20	0.01	0.06
Bike	DTR-IEV	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
	LR-R^2N	0	0.03	0	0.03	0.02	0.03	0	0	0	0	0.02	0.25	0.06	0	0.12	0.30	0.13
PTcom	DTR-IEV	0.05	0	0	0	0	0	0.05	0	0.11	0.14	0	0.48	0	0	0	0.15	0.03
	LR-R^2N	0	0.02	0.05	0.03	0	0.03	0.07	0.07	0.06	0.06	0.11	0.29	0.01	0.02	0.15	0.05	0.01
Walk	DTR-IEV	0	0	0	0	0.02	0	0	0	0	0	0.87	0.11	0	0	0	0	0
	LR-R^2N	0	0.01	0.08	0.02	0	0.03	0.03	0.02	0.17	0.17	0.19	0.05	0.05	0.10	0.07	0.01	0

Top three strongest values of dependency are highlighted in each model in the table to facilitate the comparison between the results of decision tree and linear regression models. Only coefficients of determination from **17 explanatory variables** presented in Table 47 were normalised for the comparison there. The 22 omitted explanatory variables did not have a coefficient of determination of any significant strength. See annex [E] for complete tables of coefficients of determination.

The first **decision tree diagram** from the version 3 of A-School-Filtered decision tree models is presented in Figure 24. The diagram shows the dependency of transport mode Bus on 4 explanatory variables (x210, x370, x500 and x610) including their thresholds, which the dependency rules are yielding to. The strength of the dependency is recorded in the first row in Table 47. The variable x610 has an importance of 0.72, thus it has share of 72 % on all information gain the 4 variables together contribute to the Bus decision tree model.

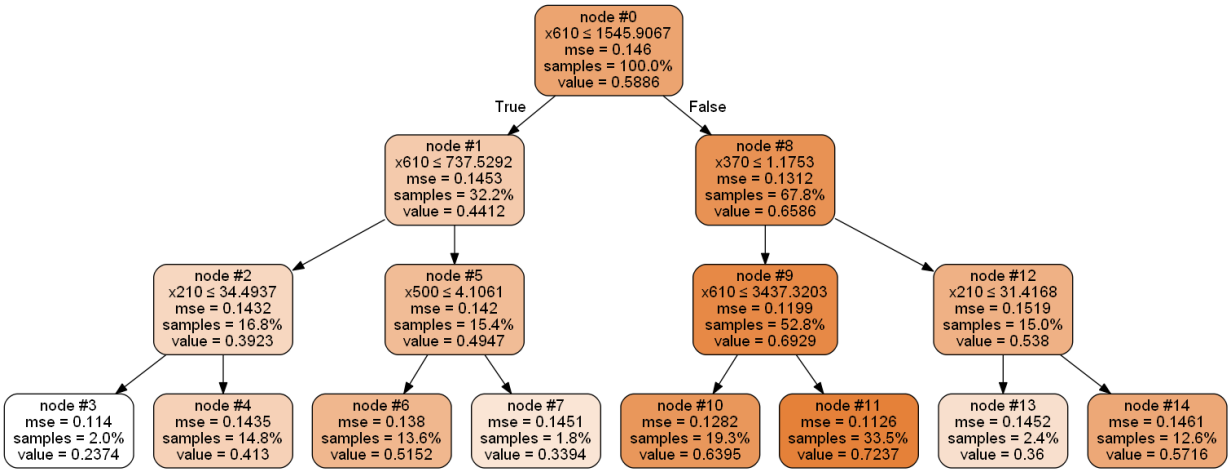


Figure 24: Decision tree diagram of transport mode Bus from its A-School-Filtered model version 3

For illustration, just the darkest terminal leaf (node #11) of the decision tree diagram with the highest value will be interpreted below. The value of leaf #11 is 0.7237, which means the average modal share of the mode Bus is 72.37 % in those 33.5 % of O-D pairs (samples), which comply with the following:

a) The distance to the closest train stop or station from the municipality of origin (explanatory variable x610) is over 3437.3203 metres
b) The ratio of inhabitants employed in key sectors (T+/PS-) in the municipality of origin (explanatory variable x370) is below or equal 1.1753
c) The condition from the root of the tree (x610 over 1545.9067 m) is already fulfilled by a)

Similarly, the other 7 terminal leaves of the decision tree diagram need to be interpreted to describe the full picture of the dependency of transport mode Bus on the explanatory variables. Decision tree diagrams of the other transport modes can be found in annex [G].

There is one important note to be made regarding the **average modal shares in leaves** of the decision tree diagrams, presented as a value of the leaf. These modal shares are calculated as an average of modal shares in all O-D pairs. This average is not weighted by the number of travellers using the transport mode in each O-D pair. On the contrary, every O-D pair has the same weight in calculation of the average modal share and thus, it is different from the modal shares calculated from the number of journeys taken by the mode and the total number of journeys, presented in chapters 5.1 and 5.2.

5.3.2 Analysis of findings

Three versions of the A-School-Filtered decision tree models were made, each with **different number of explanatory variables**. The initial version 1 used all 39 available municipality-related explanatory variables; however, some of them have not reached any importance in any of the models and were removed in the next version. The version 2 used 20 variables; again, some were removed due to unimportance. The last version 3 used only 17 explanatory variables. The difference between the numbers of used explanatory variables is significant (especially between versions 1 and 2), unlike the difference between the pruning values reviewed in Table 48. In fact, the difference in pruning values of the versions 1 through 3 is rather negligible. The minor differences might be caused either by the fact the decision tree model uses stochastic calculations, or simply because the search for local and absolute minimums of the pruning settings is influenced by a human factor.

Table 48: Comparison of pruning values of versions 1 through 3 of the A-School-Filtered decision tree models

A-School-Filtered v1 to v3		Pruned v1			Pruned v2			Pruned v3		
		RMSE			RMSE			RMSE		
Mode	**SD**	**Training**	**SD→Tr**	**Diff**	**Training**	**SD→Tr**	**Diff**	**Training**	**SD→Tr**	**Diff**
Bus	0.38152	0.35764	-6.3%	0%	0.35708	-6.4%	0%	0.35954	-5.8%	0%
Train	0.17844	0.16554	-7.2%	4%	0.16601	-7.0%	4%	0.16601	-7.0%	4%
MCT	0.10852	0.10229	-5.7%	5%	0.10072	-7.2%	6%	0.10072	-7.2%	6%
Driver	0.12287	0.12271	-0.1%	2%	0.12278	-0.1%	2%	0.12278	-0.1%	2%
CarPass	0.26991	0.26578	-1.5%	2%	0.26533	-1.7%	1%	0.26533	-1.7%	1%
PTplus	0.17018	0.17266	1.5%	8%	0.17266	1.5%	8%	0.17268	1.5%	8%
Bike	0.03926	0.04084	4.0%	23%	0.04084	4.0%	23%	0.04084	4.0%	23%
PTcom	0.17306	0.16984	-1.9%	4%	0.16951	-2.1%	4%	0.16954	-2.0%	4%
Walk	0.11938	0.11690	-2.1%	6%	0.11704	-2.0%	6%	0.11702	-2.0%	6%

Given the **minimum influence of removing the unimportant explanatory variables** to the resulting decision tree models, it is not necessary to remove these. On the other hand, if removed, the computation time can be shortened while maintaining the quality of the resulting models.

5.4 Type B models of school commute

This stage included modelling dependencies of the school commute modal shares (explained variable) on the characteristics (explanatory variables) of the **municipalities of destination** (municipalities B), using the decision tree and the linear regression models. Any of these models can be referred as a type B model, because they all use the characteristics of municipalities B as the explanatory variables.

5.4.1 Presentation of findings

The pruning table of the **B-School-Filtered decision tree models** can be seen in Table 49. As proposed in the methodology, only reduced set of 25 municipality-related explanatory variables was used in the type B models. The set was produced from the original set of 39 variables by removing the age and education related variables. As can be seen in Table 49, there is again problem with poor performance of the of PTplus and Bike models, already encountered in the A-School-Filtered models. Therefore, the problem was solved in the same way by pruning the maximum tree depth to 1.

Table 49: Pruning table – type B decision tree models of school commute using the filtered set of O-D pairs, SD = standard deviation, RMSE = root mean square error, SD→Tr = decrease from the SD to the RMSE of the training data model, Diff = difference between the RMSE resulting from training and from test data models, D = depth of the decision tree, L = number of leaves of the decision tree

B-School-Filtered		Pseudo-default						Pruned					
		RMSE				Tree		RMSE				Tree	
Mode	SD	SD→Tr	Training	Test	Diff	D	L	SD→Tr	Training	Test	Diff	D	L
Bus	0.38152	-5.7%	0.35989	0.36743	2%	6	23	-4.2%	0.36534	0.36670	0%	1	2
Train	0.17844	-3.4%	0.17234	0.16925	2%	6	42	-1.4%	0.17597	0.16846	4%	3	6
MCT	0.10852	-10.6%	0.09703	0.11015	12%	6	41	-5.6%	0.10241	0.10880	6%	4	9
Driver	0.12287	-1.7%	0.12078	0.12391	3%	6	12	0.0%	0.12292	0.12103	2%	3	4
CarPass	0.26991	-6.8%	0.25169	0.27306	8%	6	54	-2.8%	0.26230	0.26793	2%	3	6
PTplus	0.17018	-0.8%	0.16886	0.16329	3%	6	38	1.1%	0.17202	0.15824	8%	1	2
Bike	0.03926	-4.3%	0.03756	0.03928	4%	6	23	3.9%	0.04078	0.03158	23%	1	2
PTcom	0.17306	-5.7%	0.16315	0.17467	7%	6	26	-3.8%	0.16654	0.17444	5%	4	12
Walk	0.11938	-5.0%	0.11336	0.12456	9%	6	24	-2.5%	0.11640	0.12438	6%	3	6

The **strength of dependency** of modal shares of transport modes on the explanatory variables used in the B-School-Filtered models can be seen in Table 50. Both importance of explanatory variables from decision tree models (DTR-IEV) and normalised coefficients of determination from linear regression models (LR-R^2N) are presented. Out of the 25 used explanatory variables, only 18 of them actually had any importance for at least one of the decision tree models. Based on the experience with only limited change in the results of the decision tree models after removing the unimportant explanatory variables (see type A models), the type B decision tree models were not remodelled with a reduced set of explanatory variables. The decision tree diagrams can be found in annex [G].

To fit in the page, Table 50 presents only the **20 explanatory variables**, which had any importance in at least one of the decision tree models or were among the top 3 explanatory variables with the strongest dependency by the coefficient of determination. Only coefficients of determination from these 20 explanatory variables were normalised for the comparison. The 5 omitted explanatory variables did not have a coefficient of determination of any significant strength. See annex [E] for complete tables of coefficients of determination.

Table 50: Importance of the explanatory variables (DTR-IEV) and normalised coefficients of determination (LR-R^2N) in the B-School-Filtered models (using 25 explanatory variables, 20 strongest ones presented). Table 46 can be used for identification of the variables. Decision tree models of modes PTplus and Bike have poor performance and were pruned to maximum depth of 1, see Table 49. Top 1 (green), Top 2 (yellow) and Top 3 (salmon) values of dependency strength are highlighted in each row (transport mode model).

B-S-F	Model	x320	x330	x350	x360	x370	x400	x410	x420	x430	x500	x510	x600	x610	x700	x710	x720	x730	x800	x810	x820
Bus	DTR-IEV	0.06	0.03	0.02	0.08	0.20	0.05	0.07	0.11	0.01	0.09	0.03	0	0.13	0	0.05	0.04	0	0.02	0	0
	LR-R^2N	0.06	0	0	0	0.14	0	0	0.06	0.06	0.10	0.06	0	0.08	0.16	0	0.08	0	0.03	0	0
Train	DTR-IEV	0	0.10	0	0	0	0	0	0.06	0	0	0.08	0	0.77	0	0	0	0	0	0	0
	LR-R^2N	0	0.03	0	0	0	0	0.06	0.08	0.08	0.08	0.09	0.06	0.15	0	0	0.08	0.08	0	0	0
MCT	DTR-IEV	0	0	0	0.04	0.03	0	0	0	0	0	0	0	0	0.39	0.45	0	0	0	0.06	0.03
	LR-R^2N	0.08	0	0.02	0	0.17	0	0	0	0	0.11	0.08	0	0	0.25	0	0	0	0.05	0	0.01
Driver	DTR-IEV	0	0	0	0.33	0	0	0	0	0.39	0	0.27	0	0	0	0	0	0	0	0	0
	LR-R^2N	0	0.09	0	0.36	0	0.14	0.17	0	0	0	0	0	0	0	0	0.03	0.01	0	0	0
CarPass	DTR-IEV	0.18	0	0.08	0	0	0	0.09	0.54	0	0	0	0	0.11	0	0	0	0	0	0	0
	LR-R^2N	0.08	0	0.09	0.07	0	0	0.21	0.07	0.07	0.03	0.07	0	0	0	0	0	0.10	0	0	0.05
PTplus	DTR-IEV	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
	LR-R^2N	0	0	0	0	0	0	0	0.08	0.09	0.07	0.11	0.14	0.07	0	0	0.10	0.07	0	0	0
Bike	DTR-IEV	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
	LR-R^2N	0	0	0	0	0.08	0.07	0	0.09	0.10	0.06	0.10	0.09	0	0	0	0.11	0	0.06	0.08	0
PTcom	DTR-IEV	0	0	0	0.02	0.00	0.04	0.01	0.25	0.49	0	0	0	0.07	0.06	0	0	0	0.06	0	0
	LR-R^2N	0	0	0.01	0	0	0.06	0.04	0.13	0.13	0.10	0.09	0.08	0.06	0	0	0.07	0.08	0	0	0
Walk	DTR-IEV	0	0	0	0.02	0.01	0	0	0.14	0.62	0	0	0	0	0	0	0	0	0	0	0.21
	LR-R^2N	0	0.06	0	0.06	0	0	0	0.09	0.09	0	0.07	0.09	0	0	0	0.07	0.11	0	0	0.19

5.5 Type C models of school commute

This stage included modelling dependencies of the school commute modal shares (explained variable) on the characteristics (explanatory variables) of **connections** (C) between the municipalities of origin and destination, using decision tree models and linear regression models. Any of these models can be referred as a type C model, because they all use the characteristics of connections between the municipalities as the explanatory variables.

As described in the methodology (chapter 4.2.2), the set of O-D pairs used in the type C models is limited to commute relevant to administrative district of the municipality with extended powers Šternberk (ŠTBK). Only O-D pairs starting and/or ending within this district were included. In the case of school commute, that makes 53 O-D pairs. Within those 53 O-D pairs, only negligible number of journeys was taken by transport modes MCT, Bike, Walk and Rest. Moreover, the data collected within

the national population census do not allow to describe all the connections used within the journeys taken by the combined modes PTplus and PTcom. Therefore, **only connections made by transport modes Bus, Train, Driver and CarPass** were considered.

5.5.1 Presentation of findings

The pruning table of the **C-School-Filtered decision tree models** can be seen in Table 51. Only 24 connection-related explanatory variables relevant to connections by transport modes Bus, Train, Driver and CarPass were used. It was possible to effectively prune all 4 decision tree models. A conditional formatting using three colour scale is applied to SD→Tr to compare and highlight the performance of the decision tree models. The difference between the RMSE resulted from training and test data models (Diff) did not provide enough data to make an effective distinction by colour.

Table 51: Pruning table – type C decision tree models of school commute using the filtered set of O-D pairs starting and/or ending within administrative district of the municipality with extended powers Šternberk (ŠTBK), SD = standard deviation, RMSE = root mean square error, SD→Tr = decrease from the SD to the RMSE of the training data model, Diff = difference between the RMSE resulting from training and from test data models, D = depth of the decision tree, L = number of leaves of the decision tree

C-School-Filtered		Pseudo-default						Pruned					
		RMSE				Tree		RMSE				Tree	
Mode	SD	SD→Tr	Training	Test	Diff	D	L	SD→Tr	Training	Test	Diff	D	L
Bus	0.40974	89.6%	0.04249	0.17522	76%	6	14	66.9%	0.13543	0.14995	10%	2	3
Train	0.23111	100.0%	0	0.09648	N/A	3	6	53.6%	0.10733	0.11315	5%	2	3
Driver	0.13812	100.0%	0	0	N/A	4	5	24.9%	0.10371	0	N/A	2	3
CarPass	0.28021	97.5%	0.00701	0.08085	91%	6	11	38.7%	0.17183	0.2454	30%	3	4

The **strength of dependency** of the transport mode shares on the explanatory variables used in the C-School-Filtered models can be seen in Table 52. Both importance of explanatory variables from decision tree models (DTR-IEV) and normalised coefficients of determination from linear regression models (LR-R2N) are presented. Out of the 24 used explanatory variables, only 6 of them actually had any importance for at least one of the decision tree models. Based on the experience made with the type A models, the type C decision tree models were not remodelled with a reduced set of explanatory variables. The decision tree diagrams of decision tree models can be found in annex [G].

Table 52 presents the **11 explanatory variables**, which had any importance in at least one of the decision tree models or were among the top 3 explanatory variables with the strongest dependency by the coefficient of determination. Only coefficients of determination from these 11 explanatory variables were normalised for the comparison. The 13 omitted explanatory variables did not have a coefficient of determination of any significant strength. See annex [E] for complete tables of coefficients of determination.

Table 52: Importance of the explanatory variables (DTR-IEV) and normalised coefficients of determination (LR-R^2N) in the C-School-Filtered models (using 24 explanatory variables, 11 strongest ones presented). Table 13 can be used for identification of the variables. Top 1 (green), Top 2 (yellow) and Top 3 (salmon) values of dependency strength are highlighted in each row (transport mode model).

C-S-F	Model	x21	x22	x60	x70	x71	x73	x74	x80	x81	x83	x84
Bus	DTR-IEV	0	0	0	0	0	0	0	0.943	0	0	0.057
	LR-R^2N	0.018	0.003	0.052	0.282	0.060	0.000	0.015	0.442	0.089	0.006	0.033
Train	DTR-IEV	0	1E-03	0	0	0	0	0	0	0.999	0	0
	LR-R^2N	0.013	0.075	0.004	0.046	0.281	2E-04	0.003	0.068	0.508	0.002	5E-05
Driver	DTR-IEV	0	0	0	0	0	0	0	0	0	1	0
	LR-R^2N	0.201	0.138	0.008	0.008	9E-05	0.050	0.001	0.021	0.001	0.563	0.008
CarPass	DTR-IEV	0	0	7E-04	0	0	0	0	0	0	0	0.999
	LR-R^2N	0.006	4E-04	0.008	0.059	0.002	0.003	0.200	0.132	0.005	0.006	0.578

5.6 Type A models of work commute

This stage included modelling dependencies of the work commute modal shares (explained variable) on the characteristics (explanatory variables) of the **municipalities of origin** (municipalities A), using the decision tree and the linear regression models. Any of these models can be referred as a type A model, because they all use the characteristics of municipalities A as the explanatory variables.

5.6.1 Presentation of findings

The pruning table of the **A-Work-Filtered decision tree models** can be seen in Table 53. The full set of 39 municipality-related explanatory variables was used. As can be seen in the table, there is problem with a poor performance of the MCT, PTcom and Walk models, similarly to problems already encountered in models of school commute. These were again pruned maximum tree depth of 1.

Table 53: Pruning table – type A decision tree models of work commute using the filtered set of O-D pairs, SD = standard deviation, RMSE = root mean square error, SD→Tr = decrease from the SD to the RMSE of the training data model, Diff = difference between the RMSE resulting from training and from test data models, D = depth of the decision tree, L = number of leaves of the decision tree

A-Work-Filtered		Pseudo-default						Pruned					
		RMSE				Tree		RMSE				Tree	
Mode	SD	SD→Tr	Training	Test	Diff	D	L	SD→Tr	Training	Test	Diff	D	L
Bus	0.22745	-3.1%	0.22030	0.22824	3%	6	60	-1.3%	0.22438	0.22647	1%	3	5
Train	0.09552	-5.0%	0.09077	0.09395	3%	6	47	-2.6%	0.09308	0.09214	1%	3	6
MCT	0.04493	-3.0%	0.04357	0.03982	9%	6	53	2.2%	0.04594	0.03917	15%	1	2
Driver	0.34193	-3.0%	0.33184	0.34244	3%	6	64	-1.8%	0.33577	0.34081	1%	3	8
CarPass	0.18485	-1.4%	0.18232	0.18852	3%	6	54	-0.5%	0.18388	0.18732	2%	2	4
PTplus	0.14014	-2.6%	0.13646	0.14799	8%	6	42	-1.5%	0.13806	0.14712	6%	1	2
Bike	0.13950	-2.4%	0.13618	0.13676	0%	6	22	-0.6%	0.13865	0.13534	2%	2	4
PTcom	0.07051	-0.6%	0.07010	0.06937	1%	6	18	0.4%	0.07078	0.06902	2%	1	2
Walk	0.09450	-0.4%	0.09408	0.09325	1%	6	28	0.4%	0.09490	0.09278	2%	1	2

The **strength of dependency** of modal shares of transport modes on the explanatory variables used in the A-Work-Filtered models can be seen in Table 54. Both importance of explanatory variables from decision tree models (DTR-IEV) and normalised coefficients of determination from linear regression models (LR-R^2N) are presented. Out of the 39 used explanatory variables, only 11 of them actually had any importance for at least one of the decision tree models. Based on the experience made with the type A models of school commute, the work commute decision tree models were not remodelled with a reduced set of explanatory variables. The decision tree diagrams can be found in annex [G].

Table 54: Importance of the explanatory variables (DTR-IEV) and normalised coefficients of determination (LR-R^2N) in the A-Work-Filtered models (using 39 explanatory variables, 19 strongest ones presented). Table 46 can be used for identification of the variables. Decision tree models of modes MCT, PTcom and Walk have poor performance and were pruned to maximum depth of 1, see Table 53. Top 1 (green), Top 2 (yellow) and Top 3 (salmon) values of dependency strength are highlighted in each row (transport mode model).

A-W-F	Model	x220	x260	x270	x280	x300	x350	x360	x370	x410	x430	x500	x510	x610	x730	x800	x810	x820	x905	x915
Bus	DTR-IEV	0	0	0	0	0	0	0	0	0.71	0	0	0	0.29	0	0	0	0	0	0
	LR-R^2N	0.03	0.01	0.03	0.01	0.08	0.04	0.01	0.07	0.24	0	0	0.01	0.15	0	0.05	0.08	0.05	0.07	0.06
Train	DTR-IEV	0	0	0	0	0	0	0.03	0	0	0.15	0	0	0.82	0	0	0	0	0	0
	LR-R^2N	0.01	0	0.01	0	0.01	0.01	0.01	0.02	0.12	0.01	0.03	0.04	0.64	0.05	0.01	0	0.01	0	0
MCT	DTR-IEV	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
	LR-R^2N	0.06	0.07	0.06	0.05	0	0	0.01	0.15	0.01	0.10	0.17	0.17	0.02	0.07	0.01	0.02	0	0.01	0.01
Driver	DTR-IEV	0	0	0	0	0	0	0	0.12	0.83	0	0	0.06	0	0	0	0	0	0	0
	LR-R^2N	0.01	0.01	0.01	0.01	0.06	0.11	0	0.04	0.43	0	0.01	0.07	0.04	0.09	0	0.04	0.01	0.05	0.01
CarPass	DTR-IEV	0	0.54	0	0.28	0	0	0	0	0	0	0	0	0	0	0.18	0	0	0	0
	LR-R^2N	0.13	0.11	0.14	0.15	0.06	0.09	0.01	0.10	0.02	0.01	0	0.01	0	0	0.07	0.03	0.01	0.01	0.05
PTplus	DTR-IEV	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
	LR-R^2N	0	0	0	0	0.01	0.04	0	0.02	0.34	0	0	0.06	0.05	0.04	0.04	0.10	0.03	0.17	0.09
Bike	DTR-IEV	0	0	0	0	0	0	0	0	0	0	0.08	0	0	0	0.15	0.77	0	0	0
	LR-R^2N	0	0.01	0	0	0.02	0.02	0.01	0.02	0.01	0.03	0.01	0.01	0.09	0.04	0.30	0.18	0.04	0.03	0.20
PTcom	DTR-IEV	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
	LR-R^2N	0.02	0.04	0.02	0.03	0.03	0.01	0.01	0.10	0.14	0.11	0.07	0.12	0.14	0.07	0.02	0.07	0	0	0.02
Walk	DTR-IEV	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
	LR-R^2N	0.03	0.04	0.02	0.03	0	0	0.06	0.03	0	0.01	0.01	0.01	0.01	0	0.13	0.08	0.20	0.16	0.16

To fit in the page, Table 54 presents only the **19 explanatory variables**, which had any importance in at least one of the decision tree models or were among the top 3 explanatory variables with the strongest dependency by the coefficient of determination. Only coefficients of determination from these 19 explanatory variables were normalised for the comparison. The 20 omitted explanatory variables did not have a coefficient of determination of any significant strength. See annex [E] for complete tables of coefficients of determination.

5.7 Type B models of work commute

This stage included modelling dependencies of the work commute modal shares (explained variable) on the characteristics (explanatory variables) of the **municipalities of destination** (municipalities B), using the decision tree and the linear regression models. Any of these models can be referred as a type B model, because they all use the characteristics of municipalities B as the explanatory variables.

5.7.1 Presentation of findings

The pruning table of the **B-Work-Filtered decision tree models** can be seen in Table 55. As proposed in the methodology, only reduced set of 25 municipality-related explanatory variables was used in the type B models. The set was produced from the original set of 39 variables by removing the age and education related variables. As can be seen in Table 55, there is problem with a poor performance of the MCT and Walk models, already encountered in the A-Work-Filtered models. Therefore, the problem was solved in the same way by pruning the maximum tree depth to 1.

Table 55: Pruning table – type B decision tree models of work commute using the filtered set of O-D pairs, SD = standard deviation, RMSE = root mean square error, SD→Tr = decrease from the SD to the RMSE of the training data model, Diff = difference between the RMSE resulting from training and from test data models, D = depth of the decision tree, L = number of leaves of the decision tree

B-Work-Filtered		Pseudo-default						Pruned					
		RMSE				Tree		RMSE				Tree	
Mode	SD	SD→Tr	Training	Test	Diff	D	L	SD→Tr	Training	Test	Diff	D	L
Bus	0.22745	-4.2%	0.21786	0.22340	2%	6	56	-2.5%	0.22168	0.22428	1%	3	8
Train	0.09552	-3.6%	0.09213	0.09563	4%	6	36	-1.4%	0.09415	0.09333	1%	3	6
MCT	0.04493	-2.8%	0.04366	0.03976	9%	6	33	2.4%	0.04601	0.03930	15%	1	2
Driver	0.34193	-3.3%	0.33060	0.33923	3%	6	60	-2.0%	0.33518	0.33967	1%	3	6
CarPass	0.18485	-2.0%	0.18107	0.18882	4%	6	49	-0.7%	0.18362	0.18692	2%	2	4
PTplus	0.14014	-3.4%	0.13543	0.14729	8%	6	54	-2.1%	0.13726	0.14644	6%	2	4
Bike	0.13950	-2.9%	0.13545	0.13605	0%	6	56	-1.0%	0.13804	0.13476	2%	3	5
PTcom	0.07051	-2.0%	0.06907	0.07024	2%	6	37	-0.2%	0.07040	0.06871	2%	3	8
Walk	0.09450	-1.2%	0.09335	0.09452	1%	6	31	0.3%	0.09478	0.09260	2%	1	2

The **strength of dependency** of modal shares of transport modes on the explanatory variables used in the B-Work-Filtered models can be seen in Table 56. Both importance of explanatory variables from decision tree models (DTR-IEV) and normalised coefficients of determination from linear regression models (LR-R^2N) are presented. Out of the 25 used explanatory variables, only 14 of them actually had any importance for at least one of the decision tree models. Based on the experience made with the type A models of school commute, the work commute decision tree models were not remodelled with a reduced set of explanatory variables. The decision tree diagrams can be found in annex [G].

Table 56: Importance of the explanatory variables (DTR-IEV) and normalised coefficients of determination (LR-R^2N) in the B-Work-Filtered models (using 25 explanatory variables, 17 strongest ones presented). Table 46 can be used for identification of the variables. Decision tree models of modes MCT and Walk have poor performance and were pruned to maximum tree depth of 1, see Table 55. Top 1 (green), Top 2 (yellow) and Top 3 (salmon) values of dependency strength are highlighted in each row (transport mode model).

B-W-F	Model	x300	x350	x360	x370	x410	x420	x430	x500	x510	x610	x700	x710	x720	x730	x800	x810	x820
Bus	DTR-IEV	0	0	0	0	0.78	0	0	0	0	0	0	0	0	0	0	0.22	0
	LR-R^2N	0.01	0.03	0.04	0	0.29	0.07	0.07	0.05	0.15	0	0	0.03	0.04	0.08	0.02	0.11	0.01
Train	DTR-IEV	0	0	0	0	0.03	0	0	0	0	0.97	0	0	0	0	0	0	0
	LR-R^2N	0	0.01	0	0.04	0.09	0.07	0.07	0.08	0.11	0.34	0.01	0.03	0.06	0.05	0	0.01	0.01
MCT	DTR-IEV	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
	LR-R^2N	0	0.01	0	0.20	0	0.06	0.06	0.13	0.12	0.01	0.18	0.01	0.09	0.06	0.02	0.01	0.02
Driver	DTR-IEV	0	0	0.06	0	0.91	0	0	0	0.03	0	0	0	0	0	0	0	0
	LR-R^2N	0.01	0.06	0.05	0	0.38	0.04	0.04	0.05	0.13	0.05	0.01	0	0.01	0.11	0	0.06	0
CarPass	DTR-IEV	0	0	0.52	0.33	0	0	0	0	0	0	0	0.15	0	0	0	0	0
	LR-R^2N	0.02	0.03	0.10	0.17	0.02	0.02	0.02	0.05	0.09	0.04	0.08	0.13	0.07	0.03	0.10	0.02	0
PTplus	DTR-IEV	0	0	0	0	0.22	0.07	0	0	0.71	0	0	0	0	0	0	0	0
	LR-R^2N	0	0.03	0.05	0.01	0.17	0.10	0.10	0.09	0.18	0.07	0	0.02	0.05	0.10	0.01	0.04	0.01
Bike	DTR-IEV	0	0	0	0	0	0	0.21	0	0	0	0	0	0.01	0	0.12	0.67	0
	LR-R^2N	0	0	0.05	0.05	0	0.11	0.11	0.07	0.10	0	0.05	0.03	0.12	0.01	0.14	0.14	0
PTcom	DTR-IEV	0.08	0	0	0.06	0	0	0	0.57	0	0.21	0	0	0	0.09	0	0	0
	LR-R^2N	0.01	0	0.02	0.06	0.06	0.12	0.12	0.11	0.12	0.10	0.03	0.03	0.09	0.09	0.02	0.02	0
Walk	DTR-IEV	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
	LR-R^2N	0	0.09	0.06	0	0.06	0.09	0.09	0.04	0.05	0.03	0	0.03	0.06	0.10	0.06	0.02	0.22

To fit in the page, Table 56 presents only the **17 explanatory variables**, which had any importance in at least one of the decision tree models or were among the top 3 explanatory variables with the strongest dependency by the coefficient of determination. Only coefficients of determination from these 17 explanatory variables were normalised for the comparison. The 8 omitted explanatory variables did not have a coefficient of determination of any significant strength. See annex [E] for complete tables of coefficients of determination.

5.8 Type C models of work commute

This stage included modelling dependencies of the work commute modal shares (explained variable) on the characteristics (explanatory variables) of **connections** (C) between the municipalities of origin and destination, using decision tree models and linear regression models. Any of these models can be referred as a type C model, because they all use the characteristics of connections between the municipalities as the explanatory variables.

As described in the methodology (chapter 4.2.2), the set of O-D pairs used in the type C models is limited to commute relevant to administrative district of the municipality with extended powers Šternberk (ŠTBK). Only O-D pairs starting and/or ending within this district were included. In the case

of work commute, that makes 152 O-D pairs. Within those 152 O-D pairs, only negligible number of journeys was taken by transport modes MCT, Walk and Rest. Moreover, the data collected within the national population census do not allow to describe all the connections used within the journeys taken by the combined modes PTplus and PTcom. Therefore, **only connections made by transport modes Bus, Train, Driver, CarPass and Bike** were considered.

5.8.1 Presentation of findings

The pruning table of the **C-Work-Filtered decision tree models** can be seen in Table 57. Only 33 connection-related explanatory variables relevant to connections by transport modes Bus, Train, Driver, CarPass and Bike were used. It was possible to effectively prune all 5 decision tree models. A conditional formatting using three colour scale is applied to SD→Tr and Diff to compare and highlight the performance of the decision tree models.

Table 57: Pruning table – type C decision tree models of work commute using the filtered set of O-D pairs starting and/or ending within administrative district of the municipality with extended powers Šternberk (ŠTBK), SD = standard deviation, RMSE = root mean square error, SD→Tr = decrease from the SD to the RMSE of the training data model, Diff = difference between the RMSE resulting from training and from test data models, D = depth of the decision tree, L = number of leaves of the decision tree

C-Work-Filtered		Pseudo-default						Pruned					
		RMSE				Tree		RMSE				Tree	
Mode	**SD**	**SD→Tr**	**Training**	**Test**	**Diff**	**D**	**L**	**SD→Tr**	**Training**	**Test**	**Diff**	**D**	**L**
Bus	0.25201	91.4%	0.02159	0.09326	77%	6	17	53.4%	0.11752	0.11415	3%	2	4
Train	0.19776	99.3%	0.00144	0.21893	99%	6	13	57.6%	0.08378	0.10623	21%	2	3
Driver	0.35254	80.8%	0.06772	0.16934	60%	6	20	64.0%	0.12688	0.12848	1%	3	4
CarPass	0.16787	88.4%	0.01943	0.10097	81%	6	1	35.1%	0.10889	0.11869	8%	2	3
Bike	0.22478	94.6%	0.01217	0.13282	91%	6	18	68.9%	0.06989	0.11495	39%	3	5

The **strength of dependency** of modal shares of transport modes on the explanatory variables used in the C-Work-Filtered models can be seen in Table 58. Both importance of explanatory variables from decision tree models (DTR-IEV) and normalised coefficients of determination from linear regression models (LR-R2N) are presented. Out of the 33 used explanatory variables, only 7 of them actually had any importance for at least one of the decision tree models. Based on the experience made with the type A models of school commute, the work commute decision tree models were not remodelled with a reduced set of explanatory variables. The decision tree diagrams can be found in annex [G].

Table 58: Importance of the explanatory variables (DTR-IEV) and normalised coefficients of determination (LR-R^2N) in the C-Work-Filtered models (using 33 explanatory variables, 13 strongest ones presented). Table 13 can be used for identification of the variables. Top 1 (green), Top 2 (yellow) and Top 3 (salmon) values of dependency strength are highlighted in each row (transport mode model).

C-W-F	Model	x13	x22	x70	x71	x73	x74	x76	x80	x81	x83	x84	x86	x91
Bus	DTR-IEV	4E-04	0	0	0	0	0	0	1	0	0	0	0	0
	LR-R^2N	0.003	0.030	0.285	0.008	0.024	3E-04	0.008	0.524	0.021	0.055	0.003	0.015	0.023
Train	DTR-IEV	0	0	0	0	0	0	0	0	1	0	0	0	0
	LR-R^2N	0.001	0.071	0.016	0.271	0.010	0.002	0.003	0.022	0.540	0.053	0.002	0.004	0.005
Driver	DTR-IEV	0	0	0	0	0	0	0	0	0	1	0	0	0
	LR-R^2N	2E-04	0.007	0.024	0.022	0.224	0.013	0.012	0.063	0.052	0.484	0.024	0.069	0.006
CarPass	DTR-IEV	0	0	0	0	0	0	0	0	0	0	1	0	0
	LR-R^2N	4E-04	2E-04	0.004	6E-04	0.005	0.348	0.002	0.007	0.004	0.051	0.568	0.010	8E-05
Bike	DTR-IEV	0	0.011	0	0	0	0	0	0	0	0	0	0.989	0
	LR-R^2N	1E-04	0.003	0.013	0.006	0.080	0.016	0.084	0.013	0.002	0.076	0.017	0.601	0.089

Table 58 presents the **13 explanatory variables**, which had any importance in at least one of the decision tree models or were among the top 3 explanatory variables with the strongest dependency by the coefficient of determination. Only the coefficients from these 13 explanatory variables were normalised for the comparison. The 20 omitted variables did not have a coefficient of determination of any significant strength. See annex [E] for complete tables of coefficients of determination.

5.9 Discussion of findings

This chapter will discuss the findings presented and analysed in the previous chapters. The information for the discussion will be drawn across all relevant methods, models and variables.

5.9.1 Comparison of models by average strength of dependencies

This chapter will compare the **average strength of dependencies** of modal shares on explanatory variables in various combinations of model types, used data sets and used set of explanatory variables. Three types of models (A, B and C) were presented in the previous chapters. Model types A and B are municipality-related model types, as they are using the characteristics of municipalities as explanatory variables. The model type C is a connection-related model type using characteristics of connections between the municipalities as explanatory variables. The difference between the used explanatory variables in the model types is a difference from the substance of the matter.

Another difference between the municipality-related and connection-related models in this research is the **difference in the researched set of O-D pairs**. All types of models, should use the same set, however, due to high demands on data collection for the connection-related models described in chapter 4.2.2, the connection-related models only work with a sub-set of O-D pairs the municipality-related models are working with.

The municipality-related models utilise the so called **'local' set of O-D pairs**, which include filtered set O-D pairs relevant to local commute all over Czechia. The connection-related models are based on **'ŠTBK' set of O-D pairs** only relevant to local commute to, from and within the 'administrative district of the municipality with extended powers Šternberk', named after the Town of Šternberk.

Another difference between the municipality-related and connection-related models is the **prominent strength of dependency on the travel time share** based explanatory variables used in the connection-related models. These variables have great strength and influence on all transport modes within the connection-related models. No explanatory variable or group of variables in the municipality-related models have comparably strong dependencies and none of them is strongly dependant on the modal shares of all transport modes.

The average strength of dependencies in various combinations of model types, used sets of O-D pairs and used sets of explanatory variables will be compared based on an average coefficient of determination calculated based on the linear regression models. Unlike the decision tree models, the linear regression models did provide the strength of dependency for every pair 'modal share of transport mode – explanatory variable'. The strength of dependency of each pair was originally described by a correlation coefficient (r), which was then squared to obtain a **coefficient of determination** (R^2). The coefficient of determination is more suitable for further comparisons, because it can be averaged without compromising its information value.

The coefficients of determination of all 'transport mode – explanatory variable' pairs are organised in **tables coefficients of determination**, presented in annex [E]. The coefficients of determination in the tables are not normalised like the ones used for comparison with the importance of explanatory variables, presented in chapters 5.3 through 5.8.

The **average coefficient of determination** (ACD) is calculated simply as an average of all values in the table of coefficients of determination. Such average coefficient of determination describes the average strength of dependency of the shares of modelled transport modes (explained variables) and the used explanatory variables within the used O-D pairs.

In principle, Table 59 is showing a **comparison of the average coefficient of determination** in the four different combinations relevant to **school commute**. First two combinations are representing type A models (municipality-related models), both using all 39 municipality-related variables, but each using a different set of O-D pairs (local versus ŠTBK). Combinations number 3 and 4 are representing type C models (connection-related models), both using the ŠTBK set of O-D pairs, but each using different set of explanatory variables (avoiding the travel time and travel time share based in the case of

combination number 3). All four combinations presented in Table 59 include only the modal shares of transport modes Bus, Train, Driver and CarPass to facilitate the comparison.

Table 59: Comparison of average coefficient of determination (ACD) in 4 different combinations of model types, used data sets and used set of explanatory variables – the case of school commute. ŠTBK = of O-D pairs starting and/or ending within administrative district of the municipality with extended powers Šternberk, TT = travel time based explanatory variables x70 – x74, TTS = travel time share based explanatory variables x80 – x84.

#	Model	Used set of O-D pairs	Used explanatory variables	ACD
1	Type A	School Filtered local	All 39 municipality-related variables	**0.00565**
2	Type A	School Filtered ŠTBK	All 39 municipality-related variables	**0.03111**
3	Type C	School Filtered ŠTBK	16 connection-related variables without TT and TTS	**0.03590**
4	Type C	School Filtered ŠTBK	All 24 connection-related variables	**0.07980**

Table 60 is in principle providing the same comparison, only for **work commute**. Moreover, one extra transport mode (Bike) was added to all four combinations.

Table 60: Comparison of average coefficient of determination (ACD) in 4 different combinations of model types, used data sets and used set of explanatory variables – the case of work commute. ŠTBK = of O-D pairs starting and/or ending within administrative district of the municipality with extended powers Šternberk, TT = travel time based explanatory variables x70 – x76, TTS = travel time share based explanatory variables x80 – x86.

#	Model	Used set of O-D pairs	Used explanatory variables	ACD
1	Type A	Work Filtered local	All 39 municipality-related variables	**0.00121**
2	Type A	Work Filtered ŠTBK	All 39 municipality-related variables	**0.00740**
3	Type C	Work Filtered ŠTBK	23 connection-related variables without TT and TTS	**0.01391**
4	Type C	Work Filtered ŠTBK	All 33 connection-related variables	**0.05485**

The double line in Table 59 and Table 60 is dividing the **areas of comparison** elaborated below:

a) Influence of narrowing the data set of O-D pairs

The average coefficient of determination has significantly increased between the combination 1 and 2, after narrowing the local set O-D pairs to only the ŠTBK set of O-D pairs. It is more than a five-fold increase in the case of school commute, respectively six-fold in the case of work commute. The found **dependencies are, in average, much stronger within sets of O-D pairs from a narrow geographical area then within sets of O-D pairs from all over the country**.

The weaker dependencies in the local set of O-D pairs might have **two possible explanations**:

- The dependencies unexplained by the models can be accounted to some **spatial indicators** of the municipalities, which were not included among the explanatory variables,
- The **socio-psychological indicators**, which have not been used at all in this research, are strongly influenced by the location of residence of the travellers.

In any case, **this comparison indicates a strong influence of the regional localisation of the O-D pairs**.

b) Influence of adding the TT- and TTS-based explanatory variables

The average coefficient of determination has significantly increased between the combination 3 and 4, after including the travel time (TT-) and travel time share (TTS-) based explanatory variables into the models. It has increased more than two times in the case of school commute, respectively nearly four times in the case of work commute. The average strength of dependencies on the TT- and TTS-based explanatory variables accounts for more than half of the overall average dependency strength. Considering that the TT- and TTS-based explanatory variables are outnumbered by the other explanatory variables, it is clear that the **TT- and TTS-based explanatory variables are the most important for explaining the modal shares in the type C models**.

Based on the comparison of the dependency strength on the explanatory variables used in the type C models presented in chapters 5.5 and 5.8, the **modal shares have the strongest dependency on the TTS-based explanatory variables**, outperforming even the TT-based ones.

However, the great **success of TTS-based explanatory variables is being spoiled** if taking into consideration their natural similarity with the modal share. Based on definition of the travel time share, there is a perfect match of the travel time share and the modal share extreme values 0 and 1. Whenever the modal share of a mode is 0 (the mode is not used within the O-D pair) or 1 (the mode is the only one used within the O-D pair), the travel time share will also equal 0 or 1 respectively. Considering a substantial part of the modal share values in this research is either 0 or 1, it is then natural that the modal share has strong dependency on the travel time share.

c) Comparison of average dependencies between A and C type models

The average coefficients of determination should not be directly compared between the combinations 2 and 3 (or other A-C combinations), because the type A and C models are using completely different sets of explanatory variables. On the other hand, an approximate comparison can be made based on the fact the average coefficients of determination is always smaller in any type A model combination and in any type C model combination. This indicates **the connection-related models** (type C) and their explanatory variables **are better in explaining the modal shares then the municipality-related models** (represented by type A models in this comparison) and their explanatory variables.

Considering the TTS-based explanatory variables were evaluated as the most important for explaining the modal shares in the type C models and the type C models were assessed to be the best in explaining the modal shares, it means the TTS-based explanatory variables are the best explanatory variables of this research, followed by the TT-based explanatory variables. That implies **the travellers are doing mostly rational modal choices, based on the value of time.** This applies especially to commuters to work, whose average coefficient of determination has increased twice as much than in

the case of school commute after adding the TT- and TTS-based explanatory variables to the type C models. The second position of the TT-based explanatory variables is important to support to this statement as the dependency of modal share on TTS-based explanatory variables can be distorted by the natural similarity of the modal share and the travel time share discussed above.

The type B models were not included in any comparison, because the municipality-related models were already represented by the type A models. Moreover, as can be seen in the tables of coefficient of determination in annex [E], **the average coefficients of determination of type B and type A models are comparable**.

5.9.2 Found associations of explanatory variables

The large number of models presented in chapters 5.3 through 5.8 also results in a large number of found dependencies of modal shares on explanatory variables. The dependencies can be also expressed as **positive or negative associations**. The negative/positive nature of association of the modal share and the explanatory variable within the decision tree model can be read from the decision tree diagrams (listed in annex [G]). The linear regression models provide the correlation coefficients, which have positive or negative sign by a nature. It is now necessary to confront these associations found by the decision tree and in the linear regression models with associations stated in the hypothesis in chapter 3.7.1.

Table 61 is providing an **overview of associations of** 'transport mode – explanatory variables' pairs. Only the pairs, in which both the decision tree model (DTR) and the linear regression model (LR) have consistently pointed at the same top 1 explanatory variable with the highest strength of the dependency, were selected for the overview. The reason is that for most of the transport modes and their shares, the DTR and LR models are in accordance only about the top 1 explanatory variable, which contributes the most, if at all, to the explanation of the modal share dependency. For both models, the sign indicating the positive/negative nature of the found association is provided as well as the strength of the dependency. It is given by the importance of explanatory variable (IEV) in the case of DTR model and by the correlation coefficient (r) in the case of LR model. The values of r were taken from the tables of correlation coefficients presented in annex [E] and were not normalised in any way. Thus, they are not suitable for any comparison, and they mainly provide the sign of the association. The signs of associations based on the DTR and LR models are then confronted with the signs of associations expected by the hypothesis (Exp Assoc), if any such expectations exist.

Table 61: Comparison of associations expected (Exp Assoc) by the hypothesis in chapter 3.7.1 with associations found by the decision tree models (DTR) and the linear regression models (LR). IEV = importance of explanatory variable, r = correlation coefficient, TOP 1 xVAR = explanatory variable with the highest value of IEV and r in the models of the given transport mode, na = no association, + positive association, - negative association. Table 9 and Table 13 can be used for identification of the explanatory variables. Shade of grey highlights the expected associations confirmed by the found ones, shade of yellow highlights the newly found association and shade of red highlights a finding in contradiction with the hypothesis.

Model type	Transport mode	TOP 1 xVAR	Exp Assoc	DTR		LR	
				Assoc	IEV	Assoc	r
A-S-F	Bus	x610	+	+	0.717	+	0.258
	Train	x610	-	-	0.882	-	-0.25
	Driver	x430	-	+	0.624	+	0.077
	CarPass	x210	na	-	0.403	-	-0.12
	PTcom	x610	na	-	0.483	-	-0.11
	Walk	x500	+	+	0.869	+	0.103
B-S-F	Train	x610	-	-	0.769	-	-0.15
C-S-F	Bus	x80	na	+	0.943	+	0.861
	Train	x81	na	+	0.999	+	0.966
	Driver	x83	na	+	1	+	0.887
	CarPass	x84	na	+	0.999	+	0.851
A-W-F	Bus	x410	-	-	0.711	-	-0.11
	Train	x610	-	-	0.819	-	-0.16
	Driver	x410	+	+	0.828	+	0.149
	PTplus	x410	na	-	1	-	-0.06
B-W-F	Bus	x410	-	-	0.784	-	-0.17
	Train	x610	-	-	0.974	-	-0.13
	Driver	x410	+	+	0.912	+	0.160
	PTplus	x510	na	+	0.707	+	0.103
C-W-F	Bus	x80	na	+	1	+	0.938
	Train	x81	na	+	1	+	0.938
	Driver	x83	na	+	1	+	0.905
	CarPass	x84	na	+	1	+	0.890
	Bike	x86	na	+	0.989	+	0.930

The **confrontation** of the found associations with the associations stipulated in the hypothesis brings up three main types of **outcomes**:

1) **The expected associations are confirmed by the found association** – it means the associations found in this research are in line with the findings from works already reviewed in chapter [3],
2) **The found association is in a contradiction with the association expected by the hypothesis** – the one found contradiction will be discussed in chapter 5.9.2.1,
3) **There are newly found associations, where no expected associations were available** – these will be discussed in chapter 5.9.2.2.

An important note regarding the associations from the decision tree models must be added. A **deficiency** of the process used for the determination of the associations from the decision tree diagrams is that the association might not be based on the whole set of O-D pairs. The association of an explanatory variable is based on the modal share within the O-D pairs (data samples), which are in the node, where the explanatory variable is used for node splitting. Therefore, unless the association is being determined for the explanatory variable used in the root node, it is always determined based only on a part of the samples and as such, it is fully valid for this part alone (which is also the reason the importance of explanatory variables can differ from the coefficient of determination). The higher is the depth of the node with explanatory variable, the smaller is usually the number of its samples and the lesser is the general applicability of the found association. Fortunately, all the top 1 explanatory variables presented in Table 61 are located also in the root node, so the general applicability of the association is assured.

5.9.2.1 Contradiction of found and expected association

Description: The share of transport mode Driver was found to have a positive association with the Total number of workers, pupils, students and apprentices (explanatory variable x430) in the municipality of origin in the case of school commute. The total number of workers, pupils, students and apprentices (x430) is strongly correlated with the population size (x420), with correlation coefficient of 0.9998. Thus, it is possible to say that the share of transport mode Driver in school commute is increasing with the increasing population size of the municipality of origin (municipality A). However, the hypothesis expected the opposite.

Proposed explanation: First of all, the hypothesis was set for a general commute, not specifically for school commute. Moreover, the assumption was based on negative association of car use and population density, which usually goes along with the population size. In this specific case, the explanation could be that with the population size also increases the average income, which increases the affordability of the car. Practically, the greater the population size and thus grater the wealth, the bigger is the chance the parents can purchase more cars, which they can give or lend to their studying children. The higher population municipalities could also offer more working opportunities to students and apprentices to earn money for their own car.

Justification: Since the data on income of inhabitants are available for no smaller areas than the regions, it is not possible to confirm the proposed explanation by an analysis of the wealth increase in relation to the increasing population size of the municipalities.

5.9.2.2 Newly found associations

The newly found associations resulting from the municipality-related models will be discussed first, followed by the associations resulting from the connection-related models.

Negative association of CarPass with the Share of employees with apprenticeship certificate (x210)

Description: The share of transport mode CarPass was found to have a negative association with the Share of employees with apprenticeship certificate (explanatory variable x210) within the municipality of origin in the case of school commute.

Proposed explanation: Apprenticeship certificate is obtained upon completing a lower level of secondary education (practical, typically craftmanship education). Having such education is generally associated with lower-paid jobs, typically craftsman positions or with shift work, where none of the shifts is typically starting at the same time as the school education. The reason for the positive association could be based on the fact that the parents with apprenticeship certificate cannot afford to drive their children to school or are going to the work at different time than their children to school.

Justification: The lower median gross wage of workers with only apprenticeship certificate can be seen in the publicly available wage statics from CSO (2011c). The review on shift work from CSO (2012) does not provide any relation to education, but the listed employment classes CZ-ISCO with the highest share of shift work are all suitable for workers with apprenticeship certificate. The lower car ownership of the inhabitants of Czechia with only apprenticeship certificate is reported by the national mobility survey Česko v pohybu (CDV , 2020b).

Negative association of PTcom with the Distance to the closest train stop or station (x610)

Description: The share of transport mode PTcom was found to have a negative association with the Distance to the closest train stop or station (explanatory variable x610) from the municipality of origin in the case of school commute.

Proposed explanation: The decreasing Distance to the closest train stop or station is increasing the willingness of commuters to school to use combinations of public transport modes including the mode Train.

Justification: The negative association of the share of transport mode Train with the Distance to the closest train stop or station was expected by the hypothesis and then confirmed by all four municipality-related types of models used in this research, see Table 61. Thus, it is considered proven that the use of mode Train is increasing with the decreasing Distance to the closest train stop or

station. The mode Train is assumed to be a key component of the combined mode PTcom. For example, during the aggregation of transport modes used for school commute, the combinations of public transport modes including mode Train were over 63 % of all combinations aggregated into the mode PTcom. However, this ratio was valid before the reduction and filtration of the O-D pair data set. The prevailing importance of the mode Train within the combined mode PTcom can be demonstrated by the fact the mode PTcom have always followed the same trends in modal share change during reduction and filtration as the mode Train. These trends were analysed in chapters 5.1.2 and 5.2.2.

Negative association of PTplus with the Car ownership (x410)

Description: The share of transport mode PTplus was found to have a negative association with the Car ownership (explanatory variable x410) in the municipality of origin in the case of work commute.

Proposed explanation: The increasing car ownership results in increasing usage of transport mode Driver, which leads to decreasing usage of other modes, especially the public transport modes. Based on that, also the mode PTplus combining the public and private transport modes in one journey might be affected the same way, thus by a decreasing share with the increasing car ownership.

Justification: Positive association of the share of transport mode Driver with the Car ownership was expected by the hypothesis and then confirmed by both municipality-related types of models used in this research for work commute, see Table 61. The related negative association of the share of transport mode Bus with the Car ownership was also expected and confirmed. Thus, there is a solid ground to assume the mode PTplus is in the case of work commute, following the same principles as the mode Bus. However, there is no hard data evidence on the similarity of modes PTplus and Bus.

Positive association of PTplus with the Net population density (x510)

Description: The share of transport mode PTplus was found to have a positive association with the Net population density (explanatory variable x510) in the municipality of destination in the case of work commute.

Proposed explanation: Higher Net population density is generally associated with better coverage by public transport and inferior availability of parking. Both of these factors should encourage commuting to densely populated municipalities by public transport instead of passenger car-based transport modes (Driver, CarPass). For commuters travelling from municipalities with, for any reason, lower quality of public transport services, who do not want or cannot drive all the way from their municipality of origin to the densely populated municipality of destination, the mode PTplus would be a convenient option. They could use modes like Driver, CarPass or Bike to get to the suitable public

transport hub and from there continue by one of the public transport modes. Their combined journey would be an example of usage the combined transport mode PTplus and the use of such mode primarily for commute to municipalities with high Net population density would be logical.

Justification: The positive association of quality of public transport coverage and Net population density was abroad reported by De Witte et al. (2013), Limtanakool (2006) and Camagni et al. (2002), however no such evidence is available for the municipalities of Czechia. Neither is there available any hard data evidence of negative association of parking availability and Net population density.

Positive association of modal shares with the travel time based explanatory variables (x80-x86)

Description: The shares of all transport modes modelled in the connection-related models (type C models of both school and work commute) were found to have a strong, positive association with their relevant travel time based explanatory variable. For example, the modal share of Bus was found to have a strong, positive association with the Travel time share of Bus (x80), the modal share of Train with the Travel time share of Train (x81) etc.

Proposed explanation: The strong, positive association of modal shares with their travel time based explanatory variables indicates the travellers are preferring the fastest transport options. Such behaviour is rational. It means the modal choice is done rationally with the conscious intention to minimise the generalised cost of travel, in which the travel time plays a significant role.

Justification: The high importance of travel time share for explaining the modal share is indicating how beneficial is the intermodal approach in the modal split research. The travel time share based explanatory variables are the only explanatory variables included in this research, which are comparing the performance of the transport modes and they are all the top 1 explanatory variable in terms of dependency strength. They have even surpassed the travel time based explanatory variables, which are sharing the same basis (travel time), but do not offer any intermodal comparison. The intermodal approach is in line with the long-term trends in transport modelling, which has moved from previously common unimodal transport models to recently preferred multimodal transport modes, which are taking in consideration the relations between the various transport modes. Which is in line with works of for example Boyce and Williams (2016) or Kane and Behrens (2002).

[6] Application of determinants of travel behaviour

The findings from modelling of modal share dependencies **can be applied** in the decision-making process on support (or restriction) of transport modes. The support provided by the local governments to transport modes can take many different forms, for example regular operational subsidies, ad hoc incentives or planed investment in fleet or infrastructure. If knowing the general dependencies of modal split on the explanatory variables, typically the characteristics of municipalities or connections between them, it is then possible to deduce the modal share potential of transport modes in the specific municipality or O-D pair based on their known characteristics. If the dependency predicts a potential for higher or increasing share of the transport mode, then it is advisable to support or further analyse the possibilities of providing the support to such transport mode, and vice versa.

The example of application will be provided for the dependency of the share of **transport mode Train** on the **Distance to the closest train stop or station** (explanatory variable code **x610**). The transport mode Train has the strongest dependency on this variable in all 4 municipality-related models (A-School-Filtered, B-School-Filtered, A-Work-Filtered, B-Work-Filtered). The dependencies from all of them will be included in the application. Thus, the outcomes of the application can be generalised for general commute (work and school) as well as for both types of municipalities (origin and destination). That makes the dependency of modal share of transport mode Train on the Distance to the closest train stop or station the best choice for a **proof of concept** of the application.

A short note about application of the results from the connection-related models follows. The **application of results of the type C models** should only be done after a larger sample of O-D pair will be used in them, which was not the case of this research. But the conclusion of the application of connection-related models can be predicted as follows: Consider support of modes, which are the fastest (or nearly the fastest) modes within the O-D pair, or modes, which will get close to the position of the fastest mode based on the support provided.

Starting with the proof of concept, using the dependencies of modal shares of transport mode Train on the Distance to the closest train stop or station. The dependencies are described by the **information drown from the decision tree diagrams** of transport mode Train from all four types of models. The analysis will start with the decision tree diagram of transport mode Train from the **A-School-Filtered** model version 3, which is presented in Figure 25.

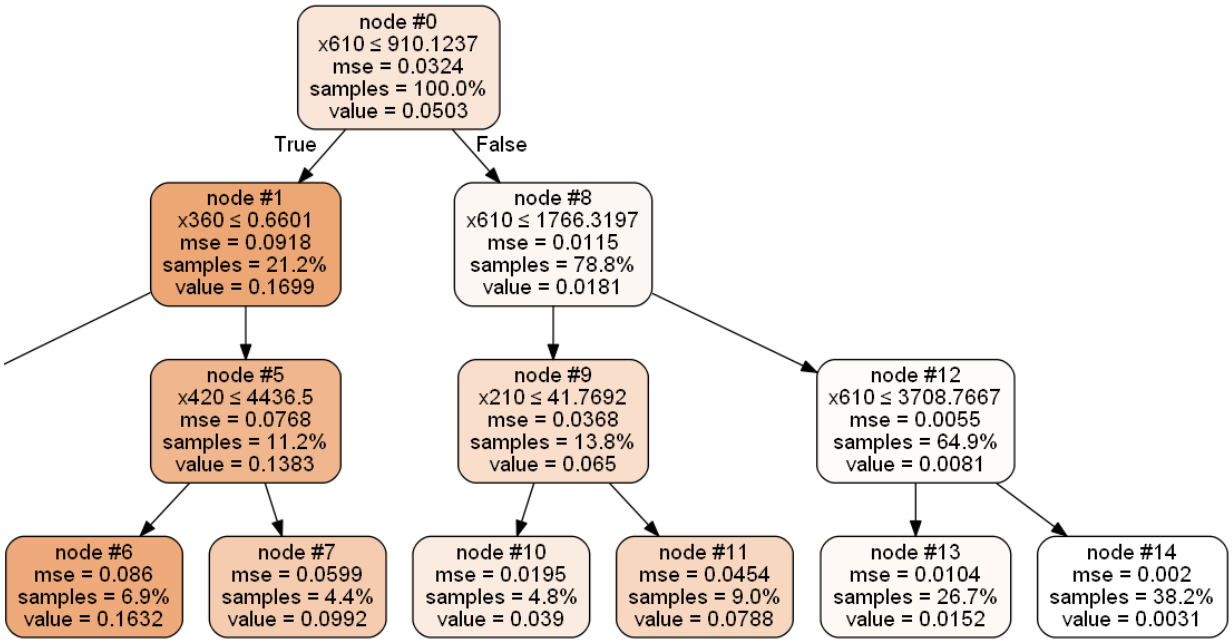


Figure 25: Decision tree diagram of transport mode Train from A-School-Filtered v3, full diagram in annex [G]

All the **information relevant to the split** of nodes of decision tree using the explanatory variable x610 were recorded in Table 62 (splitting by other variables was not considered). The table summarises the information from parent nodes, including the thresholds to be yielded in the inequality, and the information from their children nodes, mainly the fraction of samples (O-D pairs) in the node from all samples in the model, and the average modal share of the transport mode in the node (value of the node). Some of the children nodes after the split become parent nodes for another splitting. The terminal nodes of all splitting according to the explanatory variable x610 are in the table, highlighted in green colour. Those highlighted nodes together contain 100 % of all O-D pairs in the model, thus there is a ground to believe in a general applicability of the found dependency.

Table 62: Information relevant to the split of nodes using the explanatory variables Distance to the closest train stop or station (x610) in the decision tree of transport mode Train from its A-School-Filtered model version 3, terminal nodes of splitting according to the explanatory variable x610 are highlighted in green

Parent node	Variable	Threshold	Inequality yielding	Children node	Samples (O-D pairs)	Train share
#0	x610	910.12 m	below or equal	#1	21.2%	16.99%
			greater than	#8	78.8%	1.81%
#8	x610	1 766.34 m	below or equal	#9	13.8%	6.50%
			greater than	#12	64.9%	0.81%
#12	x610	3 708.77 m	below or equal	#13	26.7%	1.52%
			greater than	#14	38.2%	0.31%

The information from the 4 terminal nodes of splitting according to the explanatory variable x610 were used to create an overview of 4 **dependency intervals** of explanatory variable x610 given in Table 63. The dependency intervals are defined by the variable thresholds calculated by the decision tree model. Each interval consists of O-D pairs, which were in the terminal node, for which the interval boundaries apply. The modal share of transport mode Train within each interval is an average modal share from the respective terminal node (a value of the node). Based on the Train shares within the intervals, recommendation on support to the transport mode Train are given, using a fuzzy verbal classification.

It is important to note one thing. The **average modal shares in leaves** of the decision tree diagrams (presented as a value of the leaf): those modal shares are calculated as an average of modal shares of each O-D pair in the node. There is one modal share per each O-D pair, regardless of the number of travellers commuting within the O-D pair. Thus, the average modal share of leaf is not weighted by the number of travellers like the modal shares presented in chapters 5.1 and 5.2.

Table 63: Dependency intervals of the Train modal share on the Distance to the closest train stop or station (explanatory variable x610) defined by the variable thresholds from the A-School-Filtered decision tree model

x610 interval	Node	O-D pairs	Train share	Recommendation
below 910 m	#1	21.2%	16.99%	Definitely DO consider support
from 910 to 1 766 m	#9	13.8%	6.50%	Rather consider support
from 1 766 to 3 708 m	#13	26.7%	1.52%	Carefully consider support
over 3 708 m	#14	38.2%	0.31%	UNNECESSARY to consider support

The **fuzzy nature of the recommendations** allows the thresholds to be rounded before being used as interval boundaries of the explanatory variable. Eventually, the above boundaries could be rounded even more to obtain a simple, well rememberable 'rule of thumb', suitable for presentation to decision makers or general public. For example, in the case of school commute, the level of support to transport mode Train could be bounded by the 'Distance to the closest train stop or station from the municipality of origin' within 4 dependency intervals with boundaries of 900 – 1800 – 3600 metres, with the highest level of support for the interval of the shortest distances.

The decision tree diagrams of mode Train from **B-School-Filtered**, **A-Work-Filtered** and **B-Work-Filtered** models will follow along with the tables resulting from their analysis as well as the tables of generated dependency intervals for each model. All tables were generated the same way as the tables relevant to A-School-Filtered model presented above. Unless otherwise specified below, all highlighted nodes in each model contain 100 % of all O-D pairs of the model, thus there is a ground to believe in a general applicability of the found dependency.

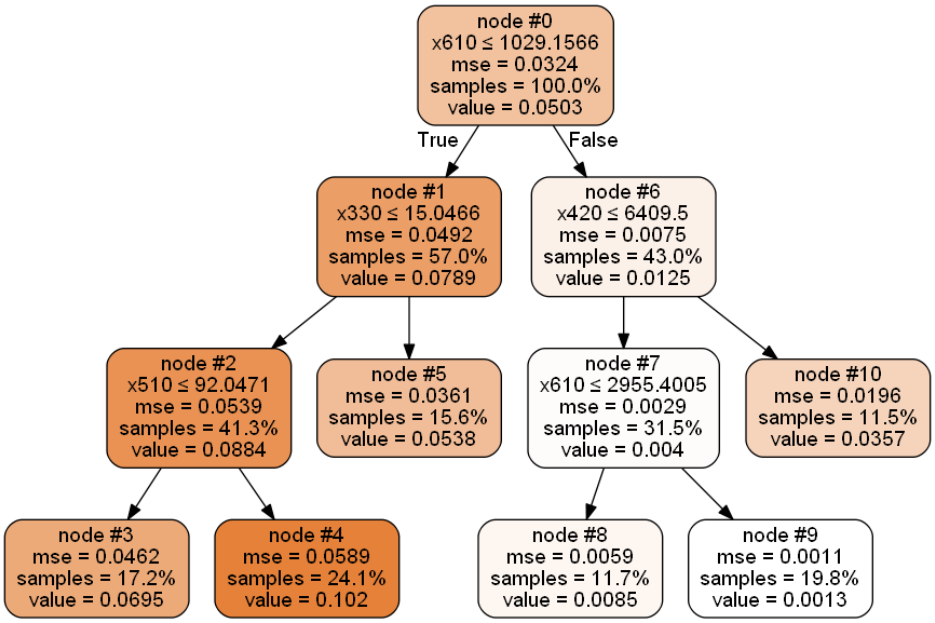


Figure 26: Decision tree diagram of transport mode Train from B-School-Filtered

All the information relevant to the split of nodes of the decision tree using the explanatory variable x610 were recorded in Table 64 based on Figure 26.

Table 64: Information relevant to the split of nodes using the explanatory variables Distance to the closest train stop or station (x610) in decision tree of transport mode Train from its B-School-Filtered model, terminal nodes of splitting according to the explanatory variable x610 are highlighted in green

Parent node	Variable	Threshold	Inequality yielding	Children node	Samples (O-D pairs)	Train share
#0	x610	1 029.16 m	below or equal	#1	57.0%	7.89%
			greater than	#6	43.0%	1.25%
#7	x610	2 955.40 m	below or equal	#8	11.7%	0.85%
			greater than	#9	19.8%	0.13%

The dependency intervals in Table 65 were generated based on the information in Table 64 highlighted in green. The highlighted terminal nodes of splitting according to the explanatory variable x610 together contain only 88.5 % of all O-D pairs in the model. Thus, there is a risk the found dependency intervals will not have a general applicability.

Table 65: Dependency intervals of the Train modal share on the Distance to the closest train stop or station (explanatory variable x610) defined by the variable thresholds from the B-School-Filtered decision tree model

x610 interval	Node	O-D pairs	Train share	Recommendation
below 1 029 m	#1	57.0%	7.89%	Definitely DO consider support
from 1 029 to 2 955 m	#8	11.7%	0.85%	Carefully consider support
over 2 955 m	#9	19.8%	0.13%	UNNECESSARY to consider support

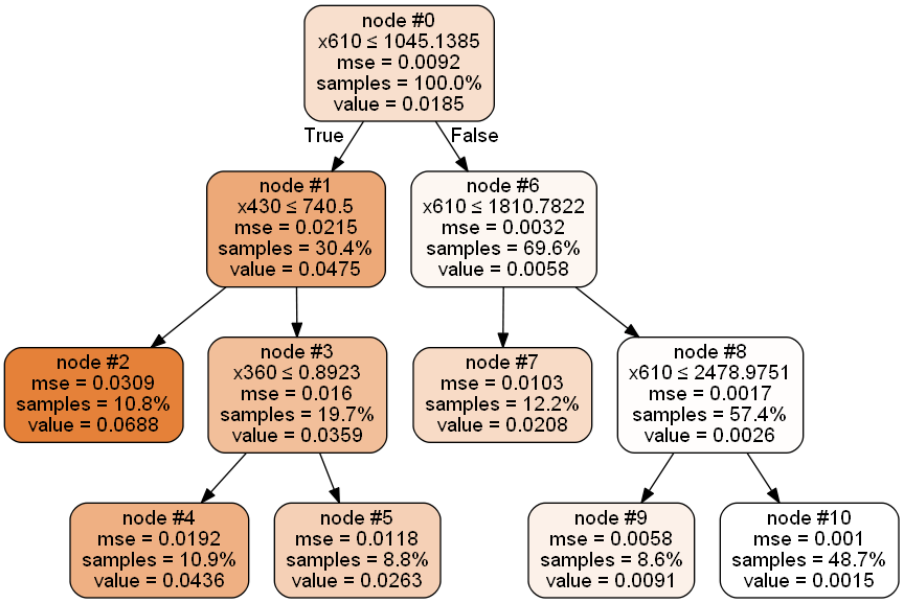


Figure 27: Decision tree diagram of transport mode Train from A-Work-Filtered

All the information relevant to the split of nodes of decision tree using the explanatory variable x610 were recorded in Table 66 based on Figure 27.

Table 66: Information relevant to the split of nodes using the explanatory variables Distance to the closest train stop or station (x610) in decision tree of transport mode Train from its A-Work-Filtered model, terminal nodes of splitting according to the explanatory variable x610 are highlighted in green

Parent node	Variable	Threshold	Inequality yielding	Children node	Samples (O-D pairs)	Train share
#0	x610	1 045.14 m	below or equal	#1	30.4%	4.75%
			greater than	#6	69.6%	0.58%
#6	x610	1 810.78 m	below or equal	#7	12.2%	2.08%
			greater than	#8	57.4%	0.26%
#8	x610	2 478.98 m	below or equal	#9	8.6%	0.91%
			greater than	#10	48.7%	0.15%

The dependency intervals in Table 67 were generated based on the information in green in Table 66.

Table 67: Dependency intervals of the Train modal share on the Distance to the closest train stop or station (explanatory variable x610) defined by the variable thresholds from the A-Work-Filtered decision tree model

x610 interval	Node	O-D pairs	Train share	Recommendation
below 1 045 m	#1	30.4%	4.8%	Definitely DO consider support
from 1 045 to 1 811 m	#7	12.2%	2.1%	Rather consider support
from 1 811 to 2 479 m	#9	8.6%	0.9%	Carefully consider support
over 2 479 m	#10	48.7%	0.2%	UNNECESSARY to consider support

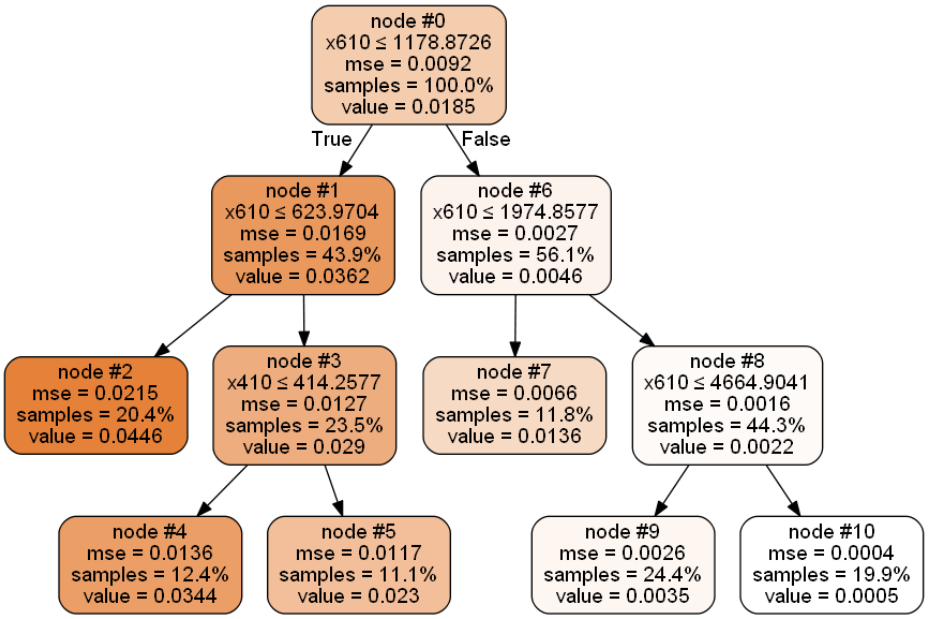


Figure 28: Decision tree diagram of transport mode Train from B-Work-Filtered

Table 68 and Table 69 were generated based on Figure 28. As Josh Starmer says: "Triple bam!!!" (2019)

Table 68: Information relevant to the split of nodes using the explanatory variables Distance to the closest train stop or station (x610) in decision tree of transport mode Train from its B-Work-Filtered model, terminal nodes of splitting according to the explanatory variable x610 are highlighted in green

Parent node	Variable	Threshold	Inequality yielding	Children node	Samples (O-D pairs)	Train share
#1	x610	623.97 m	below or equal	#2	20.4%	4.46%
			greater than	#3	23.5%	2.90%
#0	x610	1 178.87 m	below or equal	#1	43.9%	3.62%
			greater than	#6	56.1%	0.46%
#6	x610	1 974.86 m	below or equal	#7	11.8%	1.36%
			greater than	#8	44.3%	0.22%
#8	x610	4 664.90 m	below or equal	#9	24.4%	0.35%
			greater than	#10	19.9%	0.05%

Table 69: Dependency intervals of the Train modal share on the Distance to the closest train stop or station (explanatory variable x610) defined by the variable thresholds from the B-Work-Filtered decision tree model

x610 interval	Node	O-D pairs	Train share	Recommendation
below 624 m	#2	20.4%	4.46%	Definitely DO consider support
from 624 to 1 179 m	#3	23.5%	2.90%	Rather consider support
from 1 179 to 1 975 m	#7	11.8%	1.36%	Carefully consider support
from 1 975 to 4 665 m	#9	24.4%	0.35%	UNNECESSARY to consider support
over 4 655 m	#10	19.9%	0.05%	Definitely DO NOT consider support

Four sets of intervals of the Train modal share dependency on the Distance to the closest train stop or station were generated based on the decision tree diagrams from A-School-Filtered, B-School-Filtered, A-Work-Filtered, B-Work-Filtered models. The next step should be a creation of universal dependency intervals, which would comply with the sets generated for the individual models. However, the **number of dependency intervals differs** among the generated sets of dependency intervals.

Thus, a **selection process of boundaries for universal intervals** was designed and is presented in Table 70. First, all the splitting thresholds from all four models were put in the first column of the table and ordered from the smallest to largest. Then, an increase from one threshold to another was calculated as percentage. The boundaries of new universal intervals were set to be between the original thresholds, which show the greatest increase from one to another (in bold). The averages of the selected pairs of thresholds were calculated to be the values of new boundaries.

Table 70: Selection process of boundaries for universal intervals, rules of thumb for presentation of boundaries of universal intervals and application of universal intervals to the data set of Distances to the closest train stop or station from every municipality (explanatory variable x610).

Finding boundaries of new universal intervals			**Rules of thumb for presentation of boundaries**			**Application of intervals to the data set of x610**	
Thresholds [m]	**Increase [%]**	**Average [m]**	**Minutes**	**Distance [m]**	**Distance [mile]**	**Average x610 in interval [m]**	**Municipalities in interval [#]**
624						468	1 032
	46%	**767**	**7.5**	**750**	**0.5**		
910						1 080	768
	13%						
1 029							
	2%						
1 045							
	13%						
1 179							
	50%	**1 472.5**	**15**	**1 500**	**1**		
1 766						1 855	681
	3%						
1 811							
	9%						
1 975							
	26%	**2 227**	**22.5**	**2 250**	**1.5**		
2 479						3 161	1 619
	19%						
2 955							
	25%						
3 708							
	26%	**4 186.5**	**45**	**4 500**	**3**		
4 665						6 648	2 152

The calculated (raw) values of new interval boundaries (767 m, 1 472.5 m, 2 227 m and 4 186.5 m) are not exactly easy to present nor to remember. For the purpose of easier communication of the boundaries to the key stakeholders, such as political representatives, policy makers or opinion makers; the boundaries are proposed to be simplified into one of the presented sets of **rounded values** – the **rules of thumb**.

The first one represents the **walking time to the closest train stop or station** in minutes rounded to values of **7.5 – 15 – 22.5 and 45 minutes**. This rule of thumb is based on the assumption of 1-minute walking time per 100 m section. The rounding was selected to 7.5 minutes intervals, which are one of the most common and well-known time intervals of transport services.

The second and the third rule of thumb are representing the **distance to the closest train stop or station**, expressed both in metres and in miles (for stakeholders used to the imperial units). The distances expressed in the International system of units: **750 – 1 500 – 2 250 and 4 500 m** were rounded to be in line with the walking time rule of thumb and the conversion factor between them. The distances expressed in the Imperial system of units: **0.5 – 1 – 1.5 – 3 miles** was rounded to easily rememberable numbers, despite the fact the boundary of 4 186.5 m could have been rounded to 2.5 miles. On the other hand, the increase of threshold from 3 708 to 4 665 m was not so distinct compared to the increase between the pairs of thresholds above it. Thus, the boundary of 4 186.5 m is not set as conclusively as the other boundaries. Considering the fuzzy nature of the recommendations, this is an acceptable level of uncertainty in the definition of boundaries.

The recommendations on support to the transport mode Train based on the created universal dependency intervals are stated in Table 71. The reason for '**mixing up'** the result of school and work models is that the stations are serving to commute of any purpose. Similarly, the stations can be both in the municipality of origin or destination and thus the universal intervals need to be ready for this.

Table 71: Universal dependency intervals of the Train modal share on the Distance to the closest train stop or station (explanatory variable x610) defined by the new boundaries defined in Table 70, reflecting thresholds resulted from A-School-Filtered, B-School-Filtered, A-Work-Filtered and B-Work-Filtered models

x610 universal interval	**Recommendation**
below 750 m	Definitely DO consider support
from 750 to 1 500 m	Rather consider support
from 1 500 to 2 250 m	Carefully consider support
from 2 250 to 4 500 m	UNNECESSARY to consider support
over 4 500 m	Definitely DO NOT consider support

The **application** of the universal dependency intervals **to the data set of** the Distance to the closest train stop or station (explanatory variable x610) can be seen in the right two columns of Table 70. For example, an average value of x610 in the interval between 750 and 1 500 m is 1 080 m, which is based on values of 768 municipalities, whose distance to the closest train stop or station is within the interval of 750 and 1 500 m. Table 70 also shows that there are 2 152 municipalities in Czechia, whose Distance to the closest train stop or station is too long to expect any potential for higher modal share of the transport mode Train.

Of course, the applied dependencies descried by the above-mentioned recommendations do not have the ambition to be a binding set of rules in any way. But **it could be a benchmark**, an initial expert guess in the case where no other data are available. Moreover, the found dependencies are rather describing the potential of the transport mode share, which is reached on average in other municipalities or O-D pairs with the similar characteristics (explanatory variables). So, the decisions to support a transport mode must continue to be based on the knowledge of the current situation, the current modal shares. However, the support is commonly provided with the goal of increasing the modal share of the supported transport mode and that is where the applied dependency could be used for preliminary assessment, whether the aimed increase is realistic at the given characteristics of a municipality or an O-D pair. A possibility to do such assessment easily would be a great benefit to local government, which cannot afford to undertake a full-scale professional traffic engineering assessment in the case of every decision-making on the transport mode support.

For example, in the case where the current shares of some transport mode (for example Train) would be unexpectedly high in some municipality with no prerequisites for it, (for example a long distance to any train stop or station), it must be given by some **specific local circumstances**. In respect to the local circumstances, the above described recommendations should not be used to stop or deny the support to the transport mode if its modal share is already high. However, if the modal share of the mode Train is low as expected considering the Distance to the closest train stop or station, it should not be expected that any support to the mode Train will help to a significant increase of its share and no or only reduced support should be provided in line with the described recommendations.

During the decision-making process on the transport mode support, one must also take in consideration the **network character of transport services**. For example, it would not be reasonable to avoid any investment in a modernisation of one Train stop, which is far from the municipality, if the other Train stops on the same railway line are about to be modernised. The key operational parameters, such as the length and height of the platforms, should be the same on all stops on the line. On the other hand, the extent of additional investments in the passenger comfort, such as shed or

lighting, could be rationalised on the stops with low potential for Train share increase, for example based on the long distance from the closest municipality.

The applied dependencies are a predefined set of rules. As such, they are easy to be used without the need for further analysis or comprehensive knowledge. On the other hand, the fact that they are predefined for general use is limiting their ability to cover all cases, as described above. Even despite the limitations, there is a ground to believe that **the applied dependencies can be a useful tool** in the decision-making process.

[7] Conclusions and recommendations

As Josh Starmer always says: “Hooray, we’ve made it to the end of another exciting StatQuest!” (STARMER, 2019). Well, this complex thesis was not exactly a 25 minutes long excitement, but let’s hope its condensed conclusion will be.

7.1 Conclusions on dependencies

First of all, it must be said that **some dependencies of shares of transport modes on the explanatory variables were found** during the research. Stronger dependencies were found in the type C models.

A brief **recapitulation of the approach** will follow (please, skip to next paragraph for the first conclusion). The dependency of the modal share of transport modes on various explanatory variables was researched using decision tree models and linear regression models. The research has focused on the daily local commute to work or school between the municipalities of Czechia. The dependencies were researched independently for work and for school commute. The acquired data on the number of travellers, and the transport mode they have used between the municipalities of origin and destination, allowed to calculate the modal split for every origin-destination pair. The modal share of each transport mode used within each origin-destination pair acted then as the explained variable in all models. From the perspective of the used sets of explanatory variables, the models were further divided into three types A, B and C. **Type A models** used the explanatory variables related to the municipality of origin (the municipality A). The municipality-related explanatory variables included socio-demographic indicators of the municipality (e.g. population size, education of inhabitants, car ownership) and spatial indicators (e.g. population density, distance to train stop or station, mountainousness of municipality). **Type B models** were using the same municipality-related explanatory variables but related to the municipality of destination (the municipality B). Therefore, the type A and B models are collectively referred as municipality-related models. **Type C models**, so called connection-related models, were using the explanatory variables related to the connections between the municipalities of origin and destination, such as travel distance, travel time or travel time share. Due to high demands on the data collection of the connection-related explanatory variables, the connection-related models worked only with a sub-set of origin-destination pairs the municipality-related models were working with. The explanatory variables, which have achieved the strongest dependency given by an importance of explanatory variable (provided by the decision tree model) and at the same time by a normalised coefficient of determination (provided by the linear regression model), were deemed to be successful in explaining the modal share.

The most successful municipality-related explanatory variable was the Distance to the closest train stop or station (explanatory variable code **x610**). It is because this variable has shown a strong dependency with the modal share of transport mode Train in all types of the municipality-related models. There is always a negative association between the share of the mode Train and the Distance to the closest train stop or station. In other words, the **share of mode Train is increasing with the decreasing Distance to the closest train stop or station**. It applies to both cases, i.e. distance to stop or station from the municipality of origin (modelled in type A models) as well as distance from stop or station to the municipality of destination (type B models). The dependency is strong for both work and school commute. See chapters 5.9.2 and [6] for more details.

The Distance to the closest train stop or station also has a strong dependency with the share of the mode Bus in the case of type A model of school commute. There is a positive association, which means the **share of mode Bus is increasing with the increasing Distance to the closest train stop or station**. It indicates that the transport mode Bus increases its share in school commute on the expense of the mode Train with the increasing distance to stop or station from the municipality of origin.

The second most successful municipality-related explanatory variable was the **Car ownership (x410)**. It is because this variable has shown a strong dependency with multiple modal shares in both types of municipality-related models of work commute. The Car ownership has always a positive association with the share of the mode Driver and negative association with the share of the mode Bus in work commute models. In other words, **the share of the mode Driver is increasing with the increasing Car ownership, while the share of the mode Bus is decreasing**, and vice versa. It indicates that the transport mode Driver increases its share in work commute on the expense of mode Bus, with the increasing Car ownership. Moreover, the Car ownership has a negative association with the share of the transport mode PTplus. See chapter 5.9.2.2 for more details.

The remaining successful municipality-related explanatory variables will be listed in numerical order. Thus, the next one is the **Share of employees with apprenticeship certificate (x210)**. This variable has shown a strong dependency with the modal shares of the transport mode CarPass in the type A model of school commute. There is a negative association which means **the share of the mode CarPass is decreasing with the increasing Share of employees with apprenticeship certificate** in the municipality of origin of school commute. See chapter 5.9.2.2 for more details.

The explanatory variable Share of employees with apprenticeship certificate (x210) has proven to be more beneficial in explaining the modal share in the Czech environment than the variable Share of employees with elementary education (x200), which was recommended by the review (see chapter [3]) as a **benchmark** of lower level **of educational attainment**. On the other hand, although the

variable Share of employees with university degree (x260) was not labelled as a successful explanatory variable, its dependencies were the strongest of all the variables describing the higher levels of educational attainment and thus it can be recommended as the local benchmark for those.

Another successful municipality-related explanatory variable is the **Total number of workers, pupils, students and apprentices (x430)**. This variable has shown a strong dependency with the share of the mode Driver in the type A model of school commute. There is a positive association, which means **the share of the mode Driver is increasing with the increasing Total number of workers, pupils, students and apprentices** residing in the municipality of origin of school commute. However, that is in contradiction with the association expected by the hypothesis, as discussed in chapter 5.9.2.1.

The explanatory variable Total number of workers, pupils, students and apprentices (**x430**) is strongly correlated with the explanatory variable Number of inhabitants (**x420**) with a correlation coefficient of 0.9998, thus it might be necessary to use only one of them for models, which cannot handle **mutually dependent explanatory variables**. Despite the variable Total number of workers, pupils, students and apprentices is deemed successful and the Number of inhabitants not, the latter one has also achieved strong dependencies and further analysis should be done to decide on which one is the better performing, before only one is selected. This was not done in this research as both decision tree models and simple linear regression models can handle the mutually dependent explanatory variables.

Another two strongly correlated explanatory variables are the **Gross population density (x500)** and the **Net population density (x510)** with a correlation coefficient of 0.7448. Their 'competition' ended up tied as each of them was successful in one model type. Gross population density has shown a strong dependency with the modal share of transport mode Walk in the type A model of school commute. There is a positive association, which means **the share of mode Walk is increasing with the increasing Gross population density** in the municipality of origin of school commute. Net population density has shown a strong dependency with the modal share of PTplus in the type B model of work commute. There is a positive association, which means **the share of the mode PTplus is increasing with the increasing Net population density** in the municipality of destination.

The most successful connection-related explanatory variables were the travel time share based explanatory variables (x80, x81, x83, x84 and x86). It is because these variables have dominated with their dependency strength within the type C models of both work and school commute. It also makes them **the most successful explanatory variables of this research**. There is always a positive association between the modal share of transport mode and its respective travel time share. In other words, the **modal share of a transport mode is increasing with the increasing value of its travel time share**. See chapter 5.9.2 for more details.

7.2 Conclusions on methodology

First of all, it must be said that **all the used methods have proved their benefit for the research** of the modal split of the daily local commute to work or school between the municipalities of Czechia.

A brief **recapitulation** of the key travel behaviour data pre-processing methods will follow, including the reasons for their development (please, skip to next paragraph for the first conclusion). The pre-processing methods were developed to overcome the imperfections of the raw travel behaviour data set. The **aggregation** of transport modes was rationalising the 93 possible combinations of transport modes with very different level of utilisation, resulting from the methodology of the national population census. In the census, over 10 million inhabitants of Czechia were recording their usual travel behaviour on journeys to work or school, including the usually used transport modes. The 93 combinations were aggregated into 10 transport modes with more comparable level of utilisation and in a way that no commuter journeys were lost from the data set. The **transformation** of the origin-destination pairs was dealing with a structure of the raw data set, which contained the data relevant to one origin-destination pair in multiple data rows. After transformation, all the modal split and travel time data relevant to one origin-destination pair were recorded in one data row suitable as an input to modal share models. The **reduction** was addressing the natural difference in travel behaviour among the origin-destination pairs to, from and between the large towns and cities of Czechia and the origin-destination pairs to, from and between smaller towns and municipalities. The reduction was used to select only the origin-destination pairs to, from and between the small towns and municipalities relevant to the local commute. The **filtration** was dealing with the high share of origin-destination pairs with only a small number of travellers (typically only one traveller) within them. The modal split figures from such origin-destination pairs would not be calculated from a representative sample of journeys and could distort the overall modal split figures. On the other hand, the filtration also had to taken into consideration the municipal structure of Czechia, with very small municipalities, which are typically being connected only by the origin-destination pairs with the small number of travellers. Thus, only the origin-destination pairs with the small number of travellers, which were not important to the total number of in- and out-commuting travellers to and from the municipalities, were filtered out (removed) from the data set.

Two main **contributions of the travel behaviour data pre-processing methods** should be set forth:

1) **Facilitation of the modal share dependency research** by restructuring, concentrating and focusing the modal share information within the data set of explained variables. The methods thus helped to find stronger dependencies of the modal shares on the explanatory variables,
2) **Naming, describing and defining the individual steps of the data pre-processing process.** By structuring the data pre-processing process into individual steps, the purpose and

methodology of each step can be defined separately. Similarly their outcomes can be assessed separately as well their ability to meet the set objectives. Therefore, the methodology of each step can be optimised separately, which is less a complex task with a more straightforward relation of action and reaction.

The contribution number 2) is a general contribution, which **can be found useful for other researchers** dealing with similar problems or data sets. Their benefit from the named and described individual steps of the data pre-processing process can be seen in two main areas:

i) They **do not need to spend the effort on inventing the structure** of the data pre-processing steps and on defining the purpose and assessment criteria of each step,
ii) They **can individually optimise the methodology of only selected steps** according to their needs, based on their available data and other resources.

Regarding the modelling methods, both **decision tree models and linear regression models has demonstrated their applicability** for modelling dependencies of modal shares on explanatory variables by finding consensus on several strong and meaningful dependencies.

The decision tree models were not always able to find the explanatory variables, which would reliably explain the modal shares of some transport modes. Most of those modes had the least modal shares in the data set. A conclusion was made, that the trouble with the **poor performance of the decision tree models of the less used modes** proves the decision not to distinguish and model the even more rarely used transport modes such as motorbike (Moto).

The fact it was possible to find dependencies of modal shares on explanatory variables, which are relevant to municipalities and connections of municipalities proves **the basic concept of the macroanalysis of travel behaviour is viable** and can enrich the existing set of modal split research approaches and perspectives. The key advantages of the macroanalysis relying on the municipality- and connection-related explanatory variables are the following:

A) A wide range of statistical data on municipalities is **publicly available** from various sources,
B) There is a **minimum chance of breaching the personal data protection rights** by publishing the results aggregated on the level of municipalities and the connections between them,
C) Its relevance to municipalities and connections between them makes the result to be easily **interpretable and understandable** to the general public as well as the local stakeholders.

7.3 Recommendations and future steps

This chapter will mention the key recommendations on what next steps should be done in the modal split research and what are the opportunities for future discoveries.

Use more explanatory variables related to transport services in municipalities – Based on the success of the explanatory variable Distance to the closest train stop or station (x610), there is a ground to believe the municipality-related models type A and B will find stronger dependencies if more explanatory variables characterising the transport services in the municipalities are used. For example, the variables related to availability of parking facilities in municipalities, schools and workplaces could have a significant potential. Unfortunately, no data on parking were available for this research.

Involve the travel distance in the research – The dependencies of the shares of the transport modes on the travel distance could be researched and compared to the dependencies found by other authors on this matter. However, more importantly, if knowing the distance intervals, for which the transport modes are preferred, the research of dependencies could be segregated to selected distance intervals. So for example, the dependencies of Bike usage would be researched only in journeys of up to some found length limit rather than above the limit, where a high usage of mode Bike cannot be expected. Sadly, no data on travel distance were ready for this research and measuring the distance manually for every single connection within every origin-destination pair was very time demanding.

Use more detailed variables on the employment in different sectors of the economy – Instead of the three usually used explanatory variables on educational attainment (elementary, secondary, university), in total nine education-related explanatory variables were tried out in this research, finding one which was successful despite no expectation from it (apprenticeship certificate). However, this research has employed only one explanatory variable on the employment in the different sectors of the economy. It was the Ratio of inhabitants employed in key sectors (pseudo tertiary sector versus pseudo primary and secondary sectors). This variable (x370) was the strongest among the employment related explanatory variables. Considering the detail of the data on economic activity groups offered by the Czech Statistical Office, it might be a good idea to try disaggregating the Ratio of inhabitants employed in key sectors into more explanatory variables. However, this can be done only in a research including larger municipalities than this research, where not even largely aggregated sector of economy did not have any representative in small population size municipalities.

Verification of the new travel time share based explanatory variables – The explanatory variables, derived from the travel time share, were found to have the strongest dependency within the connection-related models as well as within this research as a whole. However, the strength of their

dependency might be unnaturally increasing with the increasing number of extreme values of modal share of 0 and 1, as discussed in chapter 5.9.2.2. Thus, the performance of the travel time share based explanatory variables should be tested on data sets, which avoid these extreme values. For example a data set, which does not contain the origin-destination pairs with only one traveller, which are the major source of the modal share values of 1 or 0. However, this cannot be the data sets of origin-destination pairs to or from the smallest municipalities of Czechia, which are commonly part of the origin-destination pairs with only one traveller, and which were used in this research.

Use the developed methodology for researching data from the Population and Housing Census 2021 – This research was working with the data from 2011, when the so far last national population census was conducted. A research of the nine years old data might make this thesis look outdated. The exact opposite is the truth. Besides the fact, there are no newer data available anyway, the research of the data from 2011 allowed to develop the methodology for their processing, which will be ready for the coming 2021 data. The readiness of the methodology for processing of the coming data is one of the most important contributions of this thesis. The vision is that the application of the found dependencies of the modal shares on the explanatory variables should lead to creation of a decision assisting tool for the local governments on the support on the various transport modes. Chapter [6] does not provide any complex decision-assisting tool, but an example of application, according to which the tool can be later created, based on the 2021 data.

Make connection-related models of the whole Olomouc region based on 2021 data – During the research of dates of commissioning of the rural cycle paths in Olomouc region, it was found that the main boom of new rural cycle paths has started after the 2011 population census. New cycle paths were emerging not only in the administrative district of the municipality with extended powers Šternberk, whose origin-destination pairs were studied in this research, but also in the administrative districts Olomouc, Hranice, Lipník nad Bečvou, Přerov, Prostějov, Litovel, Mohelnice, Šumperk and mostly Uničov. Thus, in the time of 2021 census, there should be much more rural cycle paths, which could be a stimulating factor for an induction of the Bike transport between the municipalities. The comparison of the modal share of transport mode Bike in 2011 and 2021 data will be more than interesting, as well as the other comparisons possible thanks to the developed methodology.

Develop methods of automated data collection of connection-related explanatory variables – The scarcity of connection-related explanatory variables given by their time demanding manual data collection was limiting the use of the connection-related models (type C models) within this research. In order to draw the benefits of the type C models in larger scale, methods of automated data collection of the connection-related explanatory variables should be developed and utilised.

7.4 Summary of the conclusions

The **thesis has fulfilled its purpose**, which was to research and describe the methods of how to process the data relevant to the modal split of local daily commute to work or school between the municipalities of Czechia. Thanks to it, also **the main goal of the thesis was achieved**, which was to find and describe the dependencies of the modal shares of transport modes on various municipality-related and connection-related explanatory variables.

The **main contributions of the thesis are** as follows:

1) **The developed methodology of data pre-processing** of the travel behaviour data set originating from the periodically recurring Population and Housing Census, mainly including:
    a. Aggregation of transport modes,
    b. Reduction of origin-destination pairs,
    c. Filtration of origin-destination pairs;

2) **Application of the found dependency of the modal share of the transport mode Train** on the Distance to the closest train stop or station. The negative association of these two variables was one of the strongest dependencies in all municipality-related models, relevant to both work and school commute, as well as to both municipalities of origin and destination. The results of the application are the universal dependency intervals valid for all four models, and the relevant recommendations on the support to mode Train presented in Table 72;

Table 72: Universal dependency intervals of the Train modal share on the Distance to the closest train stop or station (explanatory variable x610) and relevant recommendations on support to mode Train

Distance to the closest train stop or station	Potential for modal share of mode Train	Recommendations on support to mode Train
below 750 m	Very high	Definitely DO consider support
from 750 to 1 500 m	High	Rather consider support
from 1 500 to 2 250 m	Moderate	Carefully consider support
from 2 250 to 4 500 m	Low	UNNECESSARY to consider support
over 4 500 m	Very low	Definitely DO NOT consider support

3) **The developed variable 'travel time share'.** This variable is an example of an inter-modal comparison, describing the competitiveness of the travel time of each transport mode in comparison with travel times of other transport mode within the same origin-destination pair. The travel time share based explanatory variables have reached the strongest dependencies with the modal shares within the connection-related models as well as within this research as a whole. The strong, positive association of the travel time share and the modal share confirms the travellers are rationally preferring the fastest transport options.

[8] Bibliography

sklearn.tree.DecisionTreeRegressor. [Online] © 2007 - 2019, scikit-learn developers. [Cited: 17 05 2020.] https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html?highlight=decision%20tree%20regressor#sklearn.tree.DecisionTreeRegressor.

Box and Whisker Plot. [Online] Excel Easy © 2010-2020. [Cited: 11 07 2020.] https://www.excel-easy.com/examples/box-whisker-plot.html.

SPSS tutorials. [Online] © Copyright Protected 2020 by Sigma Plus Statistiek. [Cited: 03 04 2020.] https://www.spss-tutorials.com/pearson-correlation-coefficient/.

Decision Trees. [Online] © 2007 - 2019, scikit-learn developers. [Cited: 06 04 2020.] https://scikit-learn.org/stable/modules/tree.html.

StatTrek. [Online] Copyright © 2020 StatTrek.com. [Cited: 05 04 2020.] https://stattrek.com/regression/linear-regression.aspx.

Transport Planning. [Online] © 2020 Bennett, Coleman & Co. Ltd. [Cited: 01 04 2020.] https://economictimes.indiatimes.com/definition/transport-planning.

Variable Importance. [Online] © Copyright 2016-2020 H2O.ai. [Cited: 03 04 2020.] https://docs.h2o.ai/h2o/latest-stable/h2o-docs/variable-importance.html.

AFONSO, António and VENÂNCIO, Ana. 2016. The relevance of commuting zones for regional spending efficiency. *Applied Economics.* 2016, Vol. 48, No. 10, pp. 865-877. ISSN 0143-6228.

AGOVINO, Massimiliano, CERCIELLO, Massimiliano and MUSELLA, Gaetano. 2019. The good and the bad: Identifying homogeneous groups of municipalities in terms of separate waste collection determinants in Italy. *Ecological Indicators.* 2019, Vol. 98, pp. 297-309. ISSN 1470-160X.

ALONSO, William. 1964. *Location and Land Use: Toward a General Theory of Land Rent.* Cambridge (USA) : Harvard University Press, 1964. ISBN 9780674730854.

ANALYTICS VIDHYA. 2016. Analytics Vidhya. *Tree Based Algorithms: A Complete Tutorial from Scratch (in R & Python).* [Online] Analytics Vidhya, 2016. [Cited: 03 04 2020.]

ANDERSON, David R., SWEENEY, Dennis J. and WILLIAMS, Thomas A. 2002. *Essentials of Statistics for Business and Economics.* 3rd ed. Cincinnati : South-Western College Pub, 2002. ISBN 978-0324145809.

ANTIPOVA, A., WANG, F. and WILMOT, C. 2011. Urban land uses, socio-demographic attributes and commuting: a multilevel modelling approach. *Applied Geography.* 2011, Vol. 31, No. 3, pp. 1010-1018. ISSN 0143-6228.

ASAD, Firas Hasan Alwan. 2013. *City Centres: Understanding the Travel Behaviour of Residents and the Implications for Sustainable Travel.* Salford, 2013. Doctoral thesis. University of Salford, School of Computing, Science and Engineering.

ASTON, Laura, et al. 2019. Meta-analysis of the built environment and transit use relationship in different countries. *Australasian Transport Research Forum 2019 Proceedings.* Canberra : ATRF, 2019, pp. 1-20. ISSN 1447-5251 .

AXHAUSEN, K. W. 2002. A dynamic understanding of travel demand: A sketch. *Arbeitsbericht Verkehrs- und Raumplanung.* 2002, 119. Institut für Verkehrsplanung und Transportsysteme (IVT), ETH Zurich, Zurich.

AXHAUSEN, Kay W. 2007. Concepts of Travel Behavior Research. *ResearchGate.* [Online] ResearchGate GmbH, 2007. [Cited: 17 04 2020.] https://www.researchgate.net/publication/237262766_Concepts_of_Travel_Behavior_Research.

BANISTER, D. 2008. The sustainable mobility paradigm. *Transport Policy.* 2008, Vol. 15, No. 2, pp. 73-80. ISSN 0967-070X.

BERNARD, Josef. 2012. Prostorové vzorce rozvinutosti venkovských obcí Česka. *Geografie.* 2012, Vol. 117, No. 1, pp. 72-94. ISSN 2571-421X.

BHAT, C. 1998. Analysis of travel mode and departure time choice for urban shopping trips. *Transportation Research Part B: Methodological.* 1998, Vol. 32, No. 6, pp. 361-371. ISSN 0191-2615.

BHAT, C. and SINGH, S. 2000. A comprehensive daily activity-travel generation model system for workers. *Transportation Research Part A: Policy and Practice.* 2000, Vol. 34, pp. 1-22. ISSN 0965-8564.

BILER, Stanislav, et al. 2014. *Metodika aktivně-cestovního průzkumu.* Brno : Centrum dopravního výzkumu, v.v.i., MEDIAN, s.r.o., 2014. ISBN 978-80-86502-83-0.

BLAUWENS, G., DE BAERE, P. and VAN DE VOORDE, E. 2008. *Transport Economics.* 3rd ed. Antwerpen : Uitgeverij De Boeck, 2008. ISBN 9789045527048.

BONINI, Stefano and MINI, Valentina. 2018. Simple Linear Regression Analysis. *Economics Management and Policies for Global Challenges.* [Online] University of Ferrara, 2018. [Cited: 04 04 2020.] http://www.unife.it/economia/lm.economics/lectures/statistics-for-economics-business/slides/lect-4_-simple-linear-regression-model-a.

BOYCE, David E. and WILLIAMS, Huw C.W.L. 2016. *Forecasting Urban Travel: Past, Present and Future.* Cheltenham : Edward Elgar Pub, 2016. ISBN 978-1784713607.

BRID, Rajesh S. 2018. Decision trees a simple way to visualize a decision. *Medium.* [Online] Medium Corporation, 2018. [Cited: 02 04 2020.] https://medium.com/greyatom/decision-trees-a-simple-way-to-visualize-a-decision-dc506a403aeb.

BRUNN, Eric Christian. 2013. *Better Public Transit Systems: Analyzing Investments and Performance.* 2nd ed. London : Routledge, 2013. ISBN 978-0415706001.

BUTTON, Kenneth. 2010. *Transport Economics.* 3rd ed. Cheltenham : Edward Elgar Publishing Company, 2010. ISBN 978 1 84064 189 9.

BUZÁK, Jan. 2015. *Hodnocení příležitostí dálkové železniční dopravy v České republice.* Prague, 2015. Master's thesis. CTU in Prague Faculty of Transportation Sciences, Department of Transportation Systems.

CAI, Quing. 2017. *Integrating the macroscopic and microscopic traffic safety analysis using hierarchical models.* Orlando, 2017. Doctoral thesis. University of Central Florida.

CAMAGNI, R., GIBELLI, M. C. and RIGAMONTI, P. 2002. Urban mobility and urban form: the social and environmental costs of different patterns of urban expansion. *Ecological Economics.* 2002, Vol. 40, No. 2, pp. 199-216. ISSN 0921-8009.

ČARSKÝ, Jiří and ŠATRA, Petr. 2019. Use of geographical information system for finding the influence of climatic conditions on the travel behavior in various geographic areas. *19th International Multidisciplinary Scientific GeoConference SGEM 2019.* Sofia : STEF92 Technology Ltd., 2019, pp. 1063-1070. ISBN 978-619-7408-80-5.

CDV. 2020a. *Zpráva z průzkumu Česko v pohybu.* [Online] Centrum dopravního výzkumu, v.v.i., March 2020a. [Cited: 08 04 2020.] https://www.ceskovpohybu.cz/zprava/.

CDV . 2020b. *Česko v pohybu.* [Online] Centrum dopravního výzkumu, v.v.i., March 2020b. [Cited: 08 04 2020.] https://www.ceskovpohybu.cz/.

ČEKAL, Jiří. 2006. *Jihočeský kraj: Regionálně geografická analýza prostorové mobility obyvatelstva.* Brno, 2006. Doctoral thesis. Masarykova univerzita v Brně, Přírodovědecká fakulta.

CERVERO, R. 1998. *The Transit Metropolis: A Global Inquiry.* Washington, DC. : Island Press, 1998.

CHAMBERS, Michele and DINSMORE, Thomas W. 2014. *Advanced Analytics Methodologies: Driving Business Value with Analytics.* Upper Saddle River : Pearson FT Press, 2014. ISBN 978-0-13-349860-8.

CHEN, C., GONG, H. and PAASWELL, R. 2008. Role of the built environment on mode choice decisions: additional evidence on the impact of density. *Transportation.* 2008, Vol. 35, No. 3, pp. 285-299. ISSN 0049-4488.

CIRILLO, C. and AXHAUSEN, K. W. 2002. Comparing urban activity travel behaviour. *Transportation Research Board, 81th Annual Meeting, January 13–17, Washington, DC.* 2002. 27p.

COOLS, Mario, et al. 2009. Shifting towards environment-friendly modes: profiling travelers using Q-methodology. *Transportation.* 2009, Vol. 36, No. 4, pp. 437-453. ISSN 1572-9435.

CSO . 2011a. Census Forms. *Sčítání lidu, domů a bytů 2011.* [Online] Czech Statistical Office, 2011a. [Cited: 23 06 2020.] https://www.czso.cz/documents/11308/25213311/lo_anglictina.pdf/1f7adf15-325a-425b-8f2a-b43d07828ef4.

CSO . 2011b. Census Forms. *Sčítání lidu, domů a bytů 2011.* [Online] Czech Statistical Office, 2011b. [Cited: 23 06 2020.] https://www.czso.cz/documents/11308/25213311/vysvetlivky_lo_anglictina.pdf/c1d732ff-45b1-4ea2-8b99-d6a39c039881.

CSO . 2011c. Percentage of employees, time paid and average mothly gross wages by education and by sex. *Public database.* [Online] Czech Statistical Office, 2011c. [Cited: 12 07 2020.] https://vdb.czso.cz/vdbvo2/faces/en/index.jsf?page=vystup-objekt&pvo=MZD10&z=T&f=TABULKA&katalog=30852&c=v3~8__RP2011.

CSO . 2012. V práci na směny patříme k evropské špičce. *Český statistický úřad.* [Online] Czech Statistical Office, 2012. [Cited: 12 07 2020.] https://www.czso.cz/csu/czso/cesi-jsou-pati-v-eu-v-mnozstvi-prace-na-smeny.

CSO . 2013. Vyjížďka do zaměstnání a škol. *Veřejná databáze ČSÚ.* [Online] Czech Statistical Office, 2013. [Cited: 07 04 2020.]

https://vdb.czso.cz/vdbvo2/faces/cs/index.jsf?page=statistiky&filtr=G~F_M~F_Z~F_R~F_P~_S~_null_null_#katalog=30782.

CSO. 2016. Administrativní rozdělení okresu Olomouc k 1. 1. 2016. *Český statistický úřad.* [Online] Czech Statistical Office, 2016. [Cited: 13 05 2020.] https://www.czso.cz/documents/11276/17839886/Olomouc.png/fa6909a4-660f-4694-9de6-38e0ea906fe8?version=1.1&t=1478262536548.

DAVE, S. M., RAYKUNDALIYA, D. P. and SHAH, S. N. 2013. Modeling trip attributes and feasibility study of co-ordinated bus for school trips of children. *Procedia – Social and Behavioral Sciences.* 2013, Vol. 104, pp. 650-659. ISSN 1877-0428.

DE VOS, Jonas and WITLOX, Frank. 2013. Transportation policy as spatial planning tool; reducing urban sprawl by. *Journal of Transport Geography.* 2013, Vol. 33, pp. 117-125. ISSN 0966-6923.

DE WITTE, Astrid, et al. 2013. Linking modal choice to motility: A comprehensive review. *Transportation Research Part A: Policy and Practice.* 2013, Vol. 49, pp. 329-341. ISSN 0965-8564.

DHIRAJ, K. 2019. Top 5 advantages and disadvantages of Decision Tree Algorithm. *Medium.* [Online] Medium Corporation, 2019. [Cited: 02 04 2020.] https://medium.com/@dhiraj8899/top-5-advantages-and-disadvantages-of-decision-tree-algorithm-428ebd199d9a.

DORN, D. 2009. *Essays on Inequality, Spatial Interaction, and the Demand for Skills.* St. Gallen, 2009. Doctoral thesis. Graduate School of Business Administration, Economics, Law and Social Sciences.

ESCALONA-ORCAO, A. I., SÁEZ-PÉREZ, L. A. and SÁNCHEZ-VALVERDE GARCÍA, B. 2018. Location conditions for the clustering of creative activities in extra-metropolitan areas: Analysis and evidence from Spain. *Applied Geography.* 2018, Vol. 91, pp. 1-9. ISSN 0143-6228.

EUROPEAN COMMISSION. 2019. *The European Green Deal.* Brussels, 2019.

EUROSTAT. 2019. Glossary:Transport mode. *Statistics Explained.* [Online] The European Commission, 2019. [Cited: 15 05 2020.] https://ec.europa.eu/eurostat/statistics-explained/index.php/Glossary:Transport_mode.

GERMENDIA, M., URENA, J.M. and CORONADO, J.M. 2011. Long-distance trips in a sparsely populated region: The impact of high-speed infrastructures. *Journal of Transport Geography.* 2011, Vol. 19, No. 4, pp. 537-551. ISSN 0966-6923.

GEURS, K. T. and VAN WEE, B. 2004. Accessibility Evaluation of Land-use and Transport Strategies: Review and Research Directions. *Journal of Transport Geography.* 2004, Vol. 12, pp. 127-140. ISSN 0966-6923.

GILES-CORTI, Billie, et al. 2010. The co-benefits for health of investing in active transportation. *NSW Public Health Bulletin.* 2010, Vol. 21, No. 5-6, pp. 122-127. ISSN 1834-8610.

GOERLICH GISBERT, Franciosco J., CANTARINO MARTÍ, Isidro and GIELEN, Eric. 2017. Clustering cities through urban metrics analysis. *Journal of Urban Design.* 2017, Vol. 22, No. 5, pp. 689-708. ISSN 1357-4809.

HABIB, K.M.N., TIAN, Y. and ZAMAN, H. 2011. Modelling commuting mode choice with explicit consideration of carpool in the choice set formation. *Transportation.* 2011, Vol. 38, No. 4, pp. 587-604. ISSN 0049-4488.

HEADICAR, P, BANISTER, D and PHAROAH, T. 2009. *Land use and transport: Settlement patterns and the demand for travel.* Stage 2 Background Technical Report PPRO/04/07/13. London : Halcrow Group Ltd., 2009.

HEYNS, Werner and VAN JAARSVELD, Simon. 2017. Transportation modelling in practice: connecting basic theory to practice. *WIT Transactions on State-of-the-art in Science and Engineering.* 2017, Vol. 100, pp. 3-27. ISSN 1755-8336.

HOMER, Keith. 2019. Transport modelling. *Transport for West Midlands.* [Online] West Midlands Combined Authority, 2019. [Cited: 10 04 2020.] https://www.tfwm.org.uk/strategy/data-insight/transport-modelling/.

JANDOVÁ, Dominica. 2016. *Vývoj diferenciace vzdělanostní a ekonomické úrovně v obcích Středočeského kraje v letech 1980 až 2011 se zaměřením na vliv suburbanizace.* Praha, 2016. Bachelor's thesis. Univerzita Karlova v Praze, Přírodovědecká fakulta, Katedra sociální geografie a regionálního rozvoje.

JASPERS. 2014. The Use of Transport Models in Transport Planning and Project Appraisal. *Jaspers Network.* [Online] Jaspers, 2014. [Cited: 02 04 2020.] http://jaspersnetwork.org/plugins/servlet/documentRepository/displayDocumentDetails?documentId=222.

JOSSELIN, Jean-Michel, ROCABOY, Yvon and TAVÉRA, Christophe. 2009. The influence of population size on the relevance of demand or supply models for local public goods: Evidence from France. *Papers in Regional Science.* 2009, Vol. 88, No. 3, pp. 563-575.

KAIN, J. E. and LIU, Z. 2002. Efficiency and locational consequences of government transport policies and spending in Chille. [ed.] E. I. GLAESER and J. R. MEYER. *Chile: Political Economy of Urban Development.* Cambridge (USA) : John E. Harvard University Press, Kennedy School of Government, 2002, pp. 105-195.

KAJITA, Y., et al. 2004. Structural mechanism of modal choice based on the linked structure of trip purpose and transportation. *Memoirs of the Faculty of Engineering, Kyushu University.* 2004, Vol. 64, No. 1, pp. 17-33. ISSN 1880-148X.

KANE, Lisa and BEHRENS, Roger. 2002. Transport planning models - an historical and critical review. *ResearchGate.* [Online] ResearchGate GmbH, 2002. [Cited: 27 04 2020.] https://www.researchgate.net/publication/253927190_TRANSPORT_PLANNING_MODELS_-_AN_HISTORICAL_AND_CRITICAL_REVIEW.

KAUFMANN, Vincent. 2002. *Re-thinking Mobility.* Aldershot : Ashgate, 2002. ISBN 978-1138250239.

KENWORTHY, J. and LAUBE, F. 1996. Automobile dependence in cities: an international comparison of urban transport and land use patterns with implications for sustainability. *Environmental Impact Assessment Review.* 1996, Vol. 16, pp. 279-308. ISSN 0195-9255.

KII, Masanobu, HIROTA, Keiko and MINATO, Kiyoyuki. 2005. A study on modal shift potential considering public transport operation. *ResearchGate.* [Online] ResearchGate GmbH, 2005. [Cited: 21 04 2020.] https://www.researchgate.net/publication/229053305_A_study_on_modal_shift_potential_considering_public_transport_operation.

KIM, S. and ULFARSSON, G. F. 2008. Curbing automobile use for sustainable transportation: analysis of mode choice on short home-based trips. *Transportation.* 2008, Vol. 35, No. 6, pp. 723-737. ISSN 0049-4488.

KITAMURA, R. 2009. A dynamic model system of household car ownership, trip generation, and modal split: model development and simulation experiment. *Transportation.* 2009, Vol. 36, No. 6, pp. 711-732. ISSN 0049-4488.

LI, Lorraine. 2019. *Classification and Regression Analysis with Decision Trees.* [Online] Medium Corporation, 2019. [Cited: 04 04 2020.] https://towardsdatascience.com/https-medium-com-lorrli-classification-and-regression-analysis-with-decision-trees-c43cdbc58054.

LIMTANAKOOL, N., DIJST, M. and SCHWANEN, T. 2006. The influence of socioeconomic characteristics, land use and travel time considerations on mode choice for. *Journal of Transport Geography.* 2006, Vol. 14, pp. 327-341. ISSN 0966-6923.

LITMAN, T. and STEELE, R. 2013. *Land Use Impacts on Transport. How Land Use Factors affect Travel Behavior.* Victoria : Victoria Transport Policy Institute, 2013.

MACEJKA, Petr. 2014. *Vliv struktury osídlení kraje a vybraných socioekonomických ukazatelů na dopravní poptávku.* Ostrava, 2014. Doctoral thesis. Vysoká škola báňská - Technická univerzita Ostrava.

MAIER, Karel and FRANKE, Daniel. 2015. Trendy prostorové sociálně-ekonomické polarizace v Česku 2001-2011. *Sociologický časopis.* 2015, Vol. 51, No. 1, pp. 89-123. ISSN 0038-0288.

MARSHALL, S. and BANISTER, D. 2000. Travel reduction strategies: intentions and outcomes. *Transportation Research Part A: Policy and Practice.* 2000, Vol. 34, No. 5, pp. 321-338. ISSN 0965-8564.

MARTARELLI, N. J. and NAGANO, M. S. 2016. Socioeconomic Class Of Brazilian Cities For Health, Education And Employment & Income IFDM: A Clustering Data Analysis. *IEEE Latin America Transactions.* 2016, Vol. 14, No. 3, pp. 1513-1518.

MATULIN, Marko, BOŠNJAK, Ivan and ŠIMUNOVIĆ, Ljupko. 2009. *Different approaches to the modal split calculation in urban areas.* Zagreb : University of Zagreb Faculty of Transport and Traffic Sciences, 2009.

MEISTER, K., FRICK, M. and AXHAUSEN, K. 2005. A GA-based household scheduler. *Transportation.* 2005, Vol. 32, pp. 473-494. ISSN 1572-9435.

MENDENHALL, William and SINCICH, Terry. 2006. *Statistics for Engineering and the Sciences.* 5th ed. s.l. : Pearson, 2006. ISBN 978-0131877061.

MENDIOLA, Lorea, GONZÁLEZ, Pilar and CEBOLLADA, Ángel. 2014. The link between urban development and the modal split in commuting: the case of Biscay. *Journal of Transport Geography.* 2014, Vol. 37, pp. 1-9. ISSN 0966-6923.

MILIMOL, Philip, SREELATHA, T. and SOOSAN, George. 2013. Activity based travel behavioural study and mode choice modelling. *Proceedings of International Conference on Energy and Environment-2013 (ICEE 2013).* Chennai : IJIRSET, 2013, pp. 181-190. ISSN 2319-8753.

MINAL, S. and SEKHAR, Ch. R. 2014. Mode Choice Analysis: The Data, the Models and Future Ahead. *International Journal for Traffic and Transport Engineering.* 2014, Vol. 4, No. 3, pp. 269-285. ISSN 2217-5652.

MIT. 2006. Statistics for Applications: Simple Linear Regression. *MIT OpenCourseWare.* [Online] 2006. [Cited: 14 04 2020.] https://ocw.mit.edu/courses/mathematics/18-443-statistics-for-applications-fall-2006/lecture-notes/section14.pdf.

MULÍČEK, Ondřej, et al. 2013. Vymezení funkčního území Brněnské metropolitní oblasti a Jihlavské sídelní aglomerace. *Brněnská metropolitní oblast.* [Online] Magistrát města Brna, 2013. [Cited: 13 04 2020.] https://metropolitni.brno.cz/wp-content/uploads/2017/09/BMO_JSA_vymezeni_2013_web.pdf.

NOVÁK, Jakub and NETRDOVÁ, Pavlína. 2011. Prostorové vzorce sociálně-ekonomické diferenciace obcí v České republice. *Sociologický časopis.* 2011, Vol. 47, No. 4, pp. 717-744. ISSN 0038-0288.

O'FALLON, C., SULLIVAN, C. and HENSHER, D. 2004. Constraints affecting mode choices by morning car commuters. *Transport Policy.* 2004, Vol. 11, pp. 17-29. ISSN 0967-070X.

OLSZEWSKI, Piotr S. 2007. Walking as a mode of transport - a planning and policy perspective. *Prace naukowe - Budownictwo.* 2007, Vol. 146. ISSN 0137-2297.

PASTORI, Enrico, et al. 2018. *Research for TRAN Committee - Modal shift in European transport: a way forward.* Study PE 629.182. Brussles : European Parliament, Policy Department for Structural and Cohesion Policies, 2018.

PAULLEY, N. , et al. 2006. The demand for public transport: the effects of fares, quality of service, income and car ownership. *Transport Policy.* 2006, Vol. 13, No. 4, pp. 295-306. ISSN 0967-070X.

PECHER, Caroline, et al. 2013. Typology of Alpine region using spatial-pattern indicators. *Ecological Indicators.* 2013, Vol. 24, pp. 37-47. ISSN 1470-160X.

PEDREGOSA, Fabian., et al.;. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research.* 2011, Vol. 12, pp. 2825-2830. ISSN 1533-7928.

PICKUP, L. and TOWN, S. W. 1983. *A European Study of Commuting and Its Consequences.* Dublin : The European Foundation for the Improvement of Living and Working Conditions, 1983. p. 210.

PINJARI, A. R., et al. 2007. Modeling residential sorting effects to understand the impact of the built environment on commute mode choice. *Transportation.* 2007, Vol. 34, No. 5, pp. 557-573. ISSN 0049-4488.

PRIVITERA, Donatella. 2015. Towards a Competitive Sustainable City: Cycling as an Opportunity. [book auth.] Ken D. THOMAS. *Handbook of Research on Sustainable Development and Economics.* Hershey : IGI Global, 2015, pp. 20-36. ISBN 978-1466684331.

PTV. 2012. PTV Vision: VISUM 12 Fundamentals. *PTV Group.* [Online] PTV Planung Transport Verkehr AG 2012. [Cited: 03 04 2020.] https://books.google.cz/books?id=EWFoTHJJgRgC&pg=PA102&lpg=PA102&dq=PTV+Vision:+VISUM+12+Fundamentals&source=bl&ots=4-pK2lId2H&sig=ACfU3U3UEPOJ8V8x_15Z-dbshHUuj27Otg&hl=cs&sa=X&ved=2ahUKEwjXwsbQ87bpAhXJCOwKHftOCRoQ6AEwAnoECA8QAQ#v=onepage&q=PTV%20Vision%3.

PUCHER, J., BUEHLER, R. and SEINEN, M. 2011. Bicycling renaissance in North America? An update and re-appraisal of cycling trends and policies. *Transportation Research Part A: Policy and Practice.* 2011, Vol. 45, No. 6, pp. 451-476. ISSN 0965-8564.

RASCHKA, Sebastian. n.d.. What is the difference between Pearson R and Simple Linear Regression? *SebastianRaschka.* [Online] n.d. [Cited: 04 04 2020.] https://sebastianraschka.com/faq/docs/pearson-r-vs-linear-regr.html.

ŘEHÁK, S. 1984. Vliv integrace obcí na kartografické znázorňování dojížďky do zaměstnání. *Zprávy GGÚ ČSAV.* 1984, Vol. 21, No. 2, pp. 41-45. ISSN 0375-6122.

Reisender Raumplaner. 2016. Was ist der Modal Split? *Reisender Raumplaner.* [Online] YouTube, LLC, 2016. [Cited: 06 03 2020.] https://www.youtube.com/watch?v=KfzKsyhPJCc.

RIETVELD, P. 2000. Non-motorised modes in transport systems: a multimodal chain perspective for The Netherlands. *Transportation Research Part D: Transport and Environment.* 2000, Vol. 5, No. 1, pp. 31-36. ISSN 1361-9209.

RISSEL, Chris. 2009. Active travel: a climate change mitigation strategy with co-benefits for health. *NSW Public Health Bulletin.* 2009, Vol. 20, No. 1-2, pp. 10-13. ISSN 1834-8610.

RODRIGUE, Jean-Paul. 2020. *The Geography of Transport Systems.* 5th ed. New York : Routledge, 2020. ISBN 978-0-367-36463-2.

RODRIGUE, Jean-Paul, COMTOIS, Claude and SLACK, Brian. 2013. *The Geography of Transport Systems.* 3rd ed. Abingdon : Routledge, 2013. ISBN 978–0–203–37118–3.

RONAGHAN, Stacey. 2018. The Mathematics of Decision Trees, Random Forest and Feature Importance in Scikit-learn and Spark. *Towards data science.* [Online] Medium Corporation, 2018. [Cited: 02 04 2020.] https://towardsdatascience.com/the-mathematics-of-decision-trees-random-forest-and-feature-importance-in-scikit-learn-and-spark-f2861df67e3.

RUDA, Aleš and PAVLÍKOVÁ, Kateřina. 2017. Časo-prostorová dynamika vybraných demografických ukazatelů brněnské metropolitní oblasti v letech 2001 a 2011 s využitím prostorové autokorelace. *Scientific papers of the University of Pardubice. Series D, Faculty of Economics and Administration.* 2017, Vol. 24, No. 41, pp. 145-157. ISSN 1804-8048.

SABIR, M. 2011. *Weather and Travel Behaviour.* Amsterdam, 2011. Doctoral thesis. Vrije Universiteit Amsterdam, Faculty of Economics and Business Administration, Department of Spatial Economics.

SANTOS, Georgina, et al. 2013. Factors influencing modal split of commuting journeys in medium-size European cities. *Journal of Transport Geography.* 2013, Vol. 30, pp. 127-137. ISSN 0966-6923.

ŠATRA, Petr and ČARSKÝ, Jiří. 2020. Filtration and Segregation of Origin-Destination Pairs for Modal Split Research. 2020. to be published.

SCHEINER, J. 2010. Interrelations between travel mode choice and trip distance: trends in Germany 1976-2002. *Journal of Transport Geography.* 2010, Vol. 18, No. 1, pp. 75-84. ISSN 0966-6923.

SCHWANEN, T. 2002. Urban form and commuting behaviour: a cross-European perspective. *Tijdschrift voor Economische en Sociale Geografie.* 2002, Vol. 93, No. 3, pp. 336-343. ISSN 1467-9663.

ŠENK, Petr and KOUŘIL, Petr. 2014. Průzkumy dopravního chování v ČR a zahraničí. Potřebujeme národní průzkum? *Dopravní inženýrství.* 01, 2014.

SHEN, J., SAKATA, Y. and HASHIMOTO, Y. 2009. The influence of environmental deterioration and network improvement on transport modal choice. *Environmental Science and Policy.* 2009, Vol. 12, No. 3, pp. 338-346. ISSN 1462-9011.

SONG, Yena, PRESTON, John and OGILVIE, David. 2017. New walking and cycling infrastructure and modal shift in the UK: A quasi-experimental panel study. *Transportation Research Part A: Policy and Practice.* 2017, Vol. 95, pp. 320-333. ISSN 0965-8564.

STARMER, Josh. 2019. Covariance and Correlation Part 2: Pearson's Correlation. *StatQuest with Josh Starmer.* [Online] YouTube, LLC, 2019. [Cited: 18 03 2020.] https://www.youtube.com/watch?v=xZ_z8KWkhXE.

SUSILO, Y. and MAAT, E. 2007. The influence of built environment to the trends in commuting journeys in the Netherlands. *Transportation.* 2007, Vol. 34, No. 5, pp. 589-609. ISSN 1572-9435.

SZAKONYI, Petra and MAKÓ, Emese. 2014. Defining Correlation between the Modal Split of Inhabitants and Students and the Location of Housing Areas and Schools with the Analysis of Travel Plans. *Transportation Research Procedia.* 2014, Vol. 4, pp. 271-285. ISSN 2352-1465.

TOKARCZUK, Olga. 2017. *Flights.* Melbourne : Text Publishing Company, 2017. ISBN 9781925626223.

UNGVARAI, Adam. 2019. Modal Split – Different Approaches to a Common Term. *IOP Conference Series: Materials Science and Engineering.* 2019, Vol. 603, No. 4, pp. 1-9. ISSN 1757-899X.

URBANEK, Anna. 2019. Data-Driven Transport Policy in Cities: A literature Review and Implications for Future Development. *Integration as Solution for Advanced Smart Urban Transport Systems: 15th Scientific and Technical Conference "Transport Systems. Theory & Practice 2018", Selected Papers.* Springer, 2019, pp. 61-74. ISBN 978-3-319-99477-2.

VANOUTRIVE, Thomas. 2015. The modal split of cities: a workplace-based mixed modelling perspective. *Tijdschrift voor Economische en Sociale Geografie.* 2015, Vol. 106, No. 5, pp. 503-520. ISSN 1467-9663.

VASCONCELLOS, E. A. 2005. Urban change, mobility and transport in São Paulo: three decades, three cities. *Transport Policy.* 2005, Vol. 12, pp. 91-104. ISSN 0967-070X.

VYKA, Miroslav. 2010. Železniční jízdní řád 2010/2011. *Jízdní řády ČD a ČSD.* [Online] Copyright © 2020 Miroslav Vyka, 2010. [Cited: 22 03 2020.] http://www.jizdni-rady.nanadrazi.cz/index.php?page=jizdni-rady-2010-2011.

WARDMAN, Mark and TONER, Jeremy. 2020. Is generalised cost justified in travel demand analysis? *Transportation.* 2020, Vol. 47, No. 1, pp. 75–108. ISSN 1572-9435.

WHITE, P. 2009. *Public Transport: Its Planning, Management and Operation.* 5th ed. London : Routledge, 2009. ISBN 9780415445313.

WILKEN, D. 1977. European Passenger Travel Demand Analysis and Strategy Responsive Forecast. [book auth.] E. J. VISSER. *Transport decisions in an age of uncertainty.* Dordrecht : Springer, 1977, pp. 211-221. ISBN 978-94-009-9707-3.

WINGO, L. 1961. *Transportation and Urban Land.* Washington, DC. : Resources for the Future, 1961.

YE, X., PENDYALA, R. and GOTTARDI, G. 2007. An exploration of the relationship between mode choice and complexity of trip chaining patterns. *Transportation Research Part B: Methodological.* 2007, Vol. 41, No. 1, pp. 96-113. ISSN 0191-2615.

ZHANG, M. 2004. The role of land use in travel mode choice: evidence from Boston and Hong Kong. *Journal of the American Planning Association.* 2004, Vol. 70, No. 3, pp. 344-360. ISSN 0002-8991.

ZHIKHAREVICH, B. S., RUSETSKAYA, O. V. and MLADENOVIC, N. 2015. Clustering cities based on their development dynamics and Variable neigborhood search. *Electronic Notes in Discrete Mathematics.* 2015, Vol. 47, pp. 213-220.

[9] List of figures

[10] List of tables

[11] List of used software

Microsoft Office 365

ArcMap 10.5.1

Python 3.7

Mozilla Firefox 78.0.2

Adobe Acrobat Reader DC 2020.009.20074

[12] List of annexes

[A] Census forms

[B] Correlation matrices of explanatory variables

[C] Box and Whisker plots of explanatory variables

[D] Box and Whisker plots of explained variables

[E] Outcomes of linear regression models

[F] Computation script of the used decision tree regressor model

[G] Decision tree diagrams

CZECH TECHNICAL UNIVERSITY IN PRAGUE

FACULTY OF TRANSPORTATION SCIENCES

Ing. Petr Šatra

TRAVEL BEHAVIOUR AND APPLICATION OF ITS DETERMINANTS

Doctoral thesis – Annexes

2020

CONTENT

LIST OF ABBREVIATIONS

.csv	comma separated value file format
AKA	also known as
AVG	average
A-S-F	type A model of school commute using filtered set of O-D pairs (A-School-Filtered)
A-W-F	type A model of work commute using filtered set of O-D pairs (A-Work-Filtered)
B-S-F	type B model of school commute using filtered set of O-D pairs (B-School-Filtered)
B-W-F	type B model of work commute using filtered set of O-D pairs (B-Work-Filtered)
CarPass	using the passenger car for commuting as a passenger
CDV	Centrum dopravního výzkumu, v.v.i. (Transport Research Center)
CHMI	Czech Hydrometeorological Institute (Český hydrometeorologický ústav)
CSO	Czech Statistical Office (Český statistický úřad)
C-S-F	type C model of school commute using filtered set of O-D pairs (C-School-Filtered)
C-W-F	type C model of work commute using filtered set of O-D pairs (C-Work-Filtered)
Driver	using the passenger car for commuting as a driver
DTR	decision tree
GDP	gross domestic product
GIS	Geographic Information Systems
IEV	importance of explanatory variable
LR	linear regression
MCT	mass city transport
MEP	municipality with extended powers (obec s rozšířenou působností)
MTB	mountain bike
na	no association
O-D	origin-destination
PS-	aggregation of partial primary and partial secondary sectors of the economy
PT	public transport
PTcom	combination of only public transport modes
PTplus	combinations of public transport modes with private transport modes
r	correlation coefficient
R^2	coefficient of determination
RMSE	root mean square error
SD	standard deviation
ŠTBK	Town of Šternberk or the administrative district of MEP of the same name
T+	aggregation of tertiary, partial quaternary and partial quinary sectors of the economy
Tr	training data set
TT	travel time
TTS	travel time share

[A] Census forms

A.1 Census form – persons

Formulář můžete vyplnit i na www.scitani.cz

SČÍTÁNÍ LIDU, DOMŮ A BYTŮ 2011

SČÍTACÍ LIST OSOBY

Sčítání lidu, domů a bytů v roce 2011 probíhá na základě zákona č. 296/2009 Sb. Český statistický úřad Vás žádá o spolupráci formou vyplnění údajů do sčítacích formulářů, případně formou kontroly předvyplněných údajů. Ochrana individuálních a osobních údajů podle příslušných zákonů je zajištěna. Údaje budou využity výhradně ke statistickým účelům.

Vyplňuje sčítací komisař

IDENTIFIKACE																																				
										Číslo bytu					IDOB	1	0	1	4	7	5	5	4	9	2	Číslo domu	1	0	0	1	5	8	Pořadové číslo budovy			1
Okres	P	A	R	D	U	B	I	C	E																								Kód okresu	C	Z	0	5	3	2
Obec	R	Y	B	I	T	V	Í																										PSČ	5	3	3	5	4	
Část obce	R	Y	B	I	T	V	Í																										Sčítací obvod	4	8	7	0	0	2
Ulice	Š	K	O	L	N	Í																											Číslo orient.						

Vyplňuje sčítací komisař

PŘED VYPLNĚNÍM SČÍTACÍHO FORMULÁŘE SI, PROSÍM, POZORNĚ PŘEČTĚTE POKYNY K VYPLNĚNÍ A PŘILOŽENÉ VYSVĚTLIVKY

K vyplnění formuláře použijte černou nebo modrou propisovací tužku.
Z uvedených možností vyberte vždy **pouze jednu odpověď**, pokud není uvedena možnost více odpovědí.
Vybranou odpověď označte ✕ ; chybně označené pole zcela přeškrtejte ■ a označte správné ✕ . Text i číselné údaje zapisujte dle vzorů do určených vyplňovacích polí.
Chybně napsaný údaj přeškrtněte a správný napište na volné místo u příslušné otázky.

Vzory vyplnění: ✕ 0 1 2 3 4 5 6 7 8 9
A Á Ä B C Č D Ď E É Ě F G H I Í J K L Ľ M N Ň
O Ó Ö Ô P Q R Ř S Š T Ť U Ú Ů Ü V W X Y Ý Z Ž

Příjmení N O V Á K
Jméno T O M Á Š

1. Rodné číslo / → **2. Datum narození** / / → **3. Pohlaví** muž ○ žena ○
U trojmístné koncovky nechte poslední políčko prázdné. den měsíc rok

4. Státní občanství ČR ○ jiné ○ *máte-li jiné státní občanství, uveďte název státu* ↓ bez státního občanství ○

5. Bydliště v rozhodný okamžik (26. 3. 2011) *Uveďte, zda místo, kde skutečně bydlíte a kde jste členem domácnosti (bez ohledu na místo trvalého pobytu), je:*
na adrese uvedené v záhlaví formuláře ○ jinde v České republice ○ *uveďte přesnou adresu* ↓
okres
obec
č.popis. část obce
č.orient. ulice
v jiném státě *uveďte název* ○ →

6. Bydliště jeden rok před sčítáním (26. 3. 2010)
Uvedou osoby narozené do 26. 3. 2010 včetně. v obci uvedené v záhlaví formuláře ○ jinde v České republice ○ *uveďte kde* ↓
okres
obec
v jiném státě *uveďte název* ○ →

7. Bydliště matky v době narození *Uveďte, zda bydliště Vaší matky v době Vašeho narození bylo:* v obci uvedené v záhlaví formuláře ○ jinde v České republice ○ *uveďte kde* ↓
okres
obec
v jiném státě *uveďte název* ○ →

8. Rodinný stav svobodný/svobodná ○ ženatý/vdaná ○ rozvedený/rozvedená ○ vdovec/vdova ○

9. Registrované partnerství *(osob stejného pohlaví)* trvající ○ zaniklé rozhodnutím soudu ○ zaniklé úmrtím partnera/partnerky ○

10. Mateřský jazyk *Je možné uvést dva jazyky.* český ○ slovenský ○ romský ○ polský ○ německý ○ znaková řeč ○
jiný ○ *uveďte jaký* →

OTOČTE, PROSÍM, A VYPLŇTE 2. STRANU FORMULÁŘE

XXXXX

11. Národnost
Uvedení údaje není povinné.
Je možné uvést dvě národnosti.

12. Náboženská víra *Uvedení údaje není povinné. Při uvádění církve nebo náboženské společnosti registrované v ČR uveďte název dle vysvětlivek.*

věřící - hlásící se k církvi, náboženské společnosti ○ *uveďte ke které* ↓

věřící - nehlásící se k žádné církvi ani náboženské společnosti ○ bez náboženské víry ○

13. Nejvyšší ukončené vzdělání *Uvedou osoby 15leté a starší.*

bez vzdělání ○	úplné střední všeobecné (s maturitou) ○	bakalářské (Bc., BcA.) ○
neukončené základní vzdělání ○	úplné střední odborné (s maturitou) ○	magisterské (Ing., MUDr., JUDr., PhDr., Mgr. aj.) ○
základní vzdělání ○	nástavbové studium (vč. pomaturitního studia) ○	doktorské (Ph.D., ThD., DrSc., CSc.) ○
střední vč. vyučení (bez maturity) ○	vyšší odborné vzdělání (absolutorium) ○	

14. Obor vzdělání *Uvedou osoby s vyšším než základním vzděláním (viz vysvětlivky).*

15. Počet živě narozených dětí celkem *Uvedou pouze ženy 15leté a starší.*
žádné ○ jedno ○ dvě ○ více ○ → *uveďte počet dětí*

16. Počet živě narozených dětí v současném (posledním) manželství *Uvedou pouze ženy vdané, rozvedené nebo ovdovělé.*
žádné ○ jedno ○ dvě ○ více ○ → *uveďte počet dětí*

17. Ekonomická aktivita

ZAMĚSTNANÍ:	zaměstnanci, zaměstnavatelé, samostatně činní, pomáhající ○	pracující studenti a učni ○
	ženy na mateřské dovolené (28 nebo 37 týdnů), pokud před nástupem pracovaly ○	pracující důchodci ○
NEZAMĚSTNANÍ:	hledající první zaměstnání ○	ostatní nezaměstnaní ○
EKONOMICKY NEAKTIVNÍ:	nepracující důchodci ○	ostatní s vlastním zdrojem obživy ○
	žáci, studenti, učni ○	osoby v domácnosti, děti předškolního věku, ostatní závislé osoby ○

Otázky č. 18, 19 a 20 vyplňují zaměstnaní podle současného a nezaměstnaní podle posledního zaměstnání. Ekonomicky neaktivní žáci, studenti a učni pokračují otázkou č. 21. Hledající první zaměstnání a ostatní ekonomicky neaktivní již další otázky nevyplňují.

18. Zaměstnání

19. Postavení v zaměstnání
zaměstnanci ○ podnikatelé bez zaměstnanců (OSVČ) ○ pomáhající rodinní příslušníci ○
podnikatelé se zaměstnanci (zaměstnavatelé) ○ členové produkčních družstev ○

20. Odvětví ekonomické činnosti *Uveďte název podle hlavní činnosti zaměstnavatele (viz vysvětlivky).*

Otázky č. 21, 22, 23 a 24 o dojížďce/docházce do zaměstnání nebo školy vyplňují pouze zaměstnaní a žáci, studenti a učni. Pracující studenti a učni vyplňují údaje podle dojížďky/docházky do školy.

21. Místo pracoviště nebo školy na stejné adrese, jaká je v záhlaví formuláře ○ jinde v České republice ○ *uveďte kde* ↓

okres

obec

č.p./č.o. / ulice

v jiném státě ○ *uveďte název* ↓ zaměstnání bez stálého pracoviště ○

O každodenní dojížďce nebo docházce do zaměstnání nebo školy uveďte:

22. Dopravní prostředek *Uveďte dopravní prostředek/prostředky, které obvykle používáte **při jedné cestě** do zaměstnání nebo školy.*

autobus (kromě MHD) ○	automobil - spolucestující ○	motocykl ○
městská hromadná doprava ○	vlak ○	jiný ○
automobil - řidič ○	kolo ○	žádný (pouze pěšky) ○

23. Doba trvání dojížďky/docházky *Uveďte, jak dlouho Vám trvá **jedna cesta** do zaměstnání nebo školy.*
do 14 min. ○ 15 - 29 min. ○ 30 - 44 min. ○ 45 - 59 min. ○ 60 - 89 min ○ 90 a více min. ○

Uveďte, jak často dojíždíte z obce svého bydliště do obce pracoviště:

24. Frekvence dojížďky do místa pracoviště nebo školy denně ○ týdně ○ 1 - 2x měsíčně ○ jinak ○

.....
Datum, jméno a příjmení osoby, která formulář vyplnila

DĚKUJEME VÁM ZA VYPLNĚNÍ SČÍTACÍHO LISTU OSOBY

A.2 Census form – persons – census form translation

SČÍTÁNÍ LIDU, DOMŮ A BYTŮ 2011

You can also fill out the form at www.scitani.cz

CENSUS FORM-PERSONS - CENSUS FORM TRANSLATION
Please fill the information out in the original form in Czech!!!

The Population and Housing Census in 2011 is conducted on the basis of Act No. 296/2009 Coll. The Czech Statistical Office asks for your cooperation in filling out information in the census forms, or by checking the pre-completed data. Individual and personal data is protected according to the relevant laws. Data shall be used exclusively for statistical purposes.

Census officer to complete

IDENTIFICATION Dwelling no. IDOB House no. Building sequence no.

District | District code

Municipality | Post code

Part of municipality | Census district

Street | Orientation no.

Census officer to complete

PLEASE READ THE INSTRUCTIONS AND ATTACHED EXPLANATIONS CAREFULLY BEFORE FILLING OUT THE CENSUS FORM

Use a black or blue ball-point pen for filling out the form.
Always choose **one answer only** from the listed options.
Mark the selected answer ☒ ; cross out a mistake completely ■ and mark correctly ☒ . Write text and numerical data according to the examples in the designated fields. Cross out data written by mistake and write the correct data in an empty space by the relevant question.

Examples: ☒

0 1 2 3 4 5 6 7 8 9

A Á Ä B C Č D Ď E É Ě F G H I Í J K L Ľ M N Ň
O Ó Ö Ô P Q R Ř S Š T Ť U Ú Ů Ü V W X Y Ý Z Ž

Surname

First name (names)

1. Personal identification number / →

For three-digit endings, leave the last field empty.

2. Date of birth / / →
day month year

3. Sex male female

4. Country of citizenship CZ other *if you have other state citizenship, write the name of the state* ↓ no state citizenship

5. Place of residence at decisive moment (26. 3. 2011) *State whether the place where you actually live and where you are a household member (regardless of the place of permanent residence) is:*
the address listed at the top of the form elsewhere in the Czech Republic *state the exact address* ↓

district

municipality

house no. part of municipality

orientation no. street

in another state *write the name* →

6. Place of residence one year prior to the Census (26. 3. 2010) *Stated by persons born up to 26. 3. 2010 incl.* in the municipality listed at the top of the form elsewhere in the Czech Republic *state where* ↓

district

municipality

in another state *write the name* →

7. Place of residence of mother at the time of person's birth *State whether your mother's residence at the time of your birth was:* in the municipality listed at the top of the form elsewhere in the Czech Republic *state where* ↓

district

municipality

in another state *write the name* →

8. Legal marital status single married divorced widowed

9. Registered partnership *(of persons of the same sex)* enduring dissolved by decision of court ended by death of partner

10. Mother language *You can state two languages.* Czech Slovak Romany Polish German sign language

other *state what* →

PLEASE TURN OVER AND FILL OUT SECOND SIDE OF FORM

11. Ethnicity
Information is not obligatory.
You may state two ethnicities.

12. Religious belief *Information is not obligatory. State the name pursuant to the explanatory sheet for listing a church or religious organisation registered in the Czech Republic.*

believer – adheres to a church, religious organisation *state which* ↓

believer – does not adhere to any church or religious organisation

no religious faith

13. Educational attainment *To be completed by persons 15 years and older.*

no education	complete secondary, general (incl. graduation)	Bachelor's (Bc., BcA.)
incomplete primary education	complete secondary, specialist (incl. graduation)	Master's (Ing., MUDr., JUDr., PhDr., Mgr. etc.)
primary education	further education (incl. post-graduation studies)	Doctoral (Ph.D., ThD., DrSc., CSc.)
secondary, incl.vocational (no graduation)	higher specialist education (approval of work)	

14. Field of education *For persons with higher than primary education (see explanatory sheet).*

15. Total number of children born alive none one two more
Only for women 15 years and older. → *state number*

16. Number of children born alive during current or last marriage none one two more
Only for married, divorced or widowed women. → *state number*

17. Economic activity

EMPLOYED:	employees, employers, independently working, charitable		working students and apprentices
	women on maternity leave (28 or 37 weeks) if they worked prior to start		working retirees
UNEMPLOYED:	looking for first job	other unemployed	
ECONOMICALLY INACTIVE:	non-working retirees	other with own source of livelihood	
	pupils, students, apprentices	persons at home, pre-school age children, other dependent persons	

Questions 18, 19 and 20 are to be filled out by the employed according to their present job and the unemployed according to their last job.
Economically inactive pupils, students and apprentices continue to Q.21. First-job seekers and other economically inactive do not answer further questions.

18. Occupation

19. Status in employment

employee	business person without employees (self-employed)	assisting family member
business person with employees (employer)	member of production cooperative	

20. Industry (branch of economic activity) *State the name according to the main activity of the employer (see explanatory sheet).*

Questions 21, 22, 23 and 24 about commuters to work or school are to be filled out only by employees and pupils, students and apprentices.
Working students and apprentices are to answer according to the commute to school.

21. Location of place of work or school at the same address as at the top of the form elsewhere in the Czech Republic *state where* ↓

district

municipality

house no./orient. no. / street

in another state *write name* ↓

job without permanent workplace

In regard to daily commuting to work or school, state:

22. Mode of transport *State the mode of transport which you usually use for **one journey** to work or school.*

bus (other than city transport)	car - co-traveller	motorcycle
mass city transport	train	other
car - driver	bicycle	none (on foot only)

23. Time spent on daily journey *State how long **one journey** to work or school takes you.*
up to 14 min. 15 - 29 min. 30 - 44 min. 45 - 59 min. 60 - 89 min. 90 and more min.

State how often you commute from your municipality of residence to the municipality of your workplace:

24. Frequency of journey to work or school daily weekly 1 - 2 times per month other

.....
Date, first name and surname of person who filled out form

THANK YOU FOR FILLING OUT THE CENSUS FORM-PERSONS

A.3 Explanation for the census form – persons

SČÍTÁNÍ LIDU, DOMŮ A BYTŮ 2011

EXPLANATIONS FOR THE CENSUS FORM - PERSONS

The Census Form – Persons is filled out by all people present on census day in the Czech Republic regardless of whether they live in a dwelling, boarding facilities or in another building. In counting people in a dwelling or in facilities designated for long-term residence, the form must also be filled out for persons temporarily absent if they really do live in the flat or boarding facilities and they are a member of a specific household there. For minors or people not competent to enter into legal acts, information will be provided by their legal representative. Information is filled out according to the situation at midnight from 25 to 26 March 2011, unless stated otherwise.
In questions 4,10,11 and 22 it is possible to provide more answers. For other questions, always indicate or write one answer only. Foreigners living short-term in the Czech Republic (generally up to 90 days) fill out their first name, surname and questions 2-6 only.
Write all text legibly in large printed letters according to the examples in the form. Write numerical information from the right (apart from birth numbers).

4. COUNTRY OF CITIZENSHIP
In the case of dual citizenship, state both.

5. PLACE OF RESIDENCE AT DECISIVE MOMENT (26.3.2011)
State whether the address at the top of the form (in the IDENTIFICATION field) is the place where the polled person has in fact lived long-term and where he/she has a household or family. It is not important where the person is registered for permanent residence or whether, for example, for reasons of work or study he/she spends the greater part of the week in another place. In the case that the residence address is different from the address at the top of the form, enter it as precisely as possible.

7. PLACE OF RESIDENCE OF MOTHER AT THE TIME OF PERSON´S BIRTH
Place of residence of mother at the time of person´s birth means the actual residence of the mother at the time of the birth of the polled person, i.e. the first residence where the polled person lived immediately after the birth. The information cannot be exchanged for „place of birth", which is stated, for example, in the birth certificate and indicates the actual place of birth - generally the address of the maternity hospital. State the municipality, district, or state according to the name and borders valid on census day.

8. LEGAL MARITAL STATUS
This ascertains de jure family status, or legal status. Enter a common-law partnership on the second side of the census form - dwellings only.
Each person fills out either question 8 - family status, or question 9 - registered partnership.

9. REGISTERED PARTNERSHIP
This ascertains the de jure status, or legal status - of a registered partnership of persons of the same sex entered into according to Act No. 115/2006 Coll., on Registered Partnerships, or pursuant to the valid legal regulations abroad. Persons who never entered into a registered partnership do not answer the question.

10. MOTHER LANGUAGE
The mother language should be stated as the language in which the polled person was spoken to in childhood by his/her mother or the persons who brought him/her up. It is permissible to state more languages.

11. ETHNICITY
Information about ethnicity is stated by each person according to their own decision. Stating this information is not obligatory. It is permissible to state more ethnicities. Ethnicity means affiliation to a nation, a nationality or an ethnic minority. Mother tongue or a language which the respondent primarily uses or manages best is not a determinant of ethnicity.

12. RELIGIOUS BELIEF
Information about religious belief is filled out by each person according to their own decision. Stating this information is not obligatory. In the case of listing a registered church or religious society, write the exact or a suitably shortened title according to the following overview:
Churches and religious societies registered in the Czech Republic as of 2.11.2010:
- *Apostolic Church*
- *Unity of the Brethren Baptists*
- *Diamond Way Bhuddism of the Karma Kagyu Lineage*
- *Church of the Seventh Day Adventists*
- *Brethren Church*
- *Czechoslovak Hussite Church*
- *Church of Jesus Christ of the Latter Day Saints in the Czech Republic*
- *Christian Society Church*
- *New Hope Church*
- *Greek Catholic Church*
- *Roman Catholic Church*
- *Word of Life Church*
- *Church of the Living God*
- *Czech Hindu Religious Society*
- *Evangelical Church of Czech Brethren*
- *Evangelical Church of the Augsburg Confess/on in the Czech Republic*
- *Evangelical Methodist Church*
- *Federation of Jewish Communities in the Czech Republic*
- *United Brethren*
- *Christian Fellowship*
- *Lutheran Evangelical Church of the Augsburg Confession in the Czech Republic*
- *International Society for Krishna Consciousness, Hare Krishna Movement*
- *Religious Society of Czech Unitarians*
- *Religious Society of Jehova's Witnesses*
- *New Apostolic Church in the Czech Republic*
- *Christian Community in the Czech Republic*
- *Orthodox Church in Czech Lands*
- *Russian Orthodox Church, jurisdiction of the Patriarch of Moscow and all Russia in the Czech Republic*
- *Silesian Evangelical Church of the Augsburg Confession*
- *Old Catholic Church in the Czech Republic*
- *Centre of Muslim Communities*
- *Vishva Nirmala Dharma*

If you state another church, religious society, confession or faith, write it as precisely as possible.

13. EDUCATIONAL ATTAINMENT
To be filled out by persons 15 years of age and older only, on the basis of the highest completed school.
Persons who obtained their education in a foreign educational institution or private educational institution not accredited by the Ministry of Education, Youth and Sports, should state the corresponding level of education in the Czech educational system if their education is recognised as having parity.
No education is listed by persons who did not complete even the first level of primary school.
Incomplete primary education is listed by persons who completed only the first level of primary school or the former elementary school or national school, or matriculated only at an auxilliary school.
Primary education is also listed by students of six and eight-year grammar schools, or eight-year departments of a conservatory, if they had already graduated from the first two, or four years of the relevant grammar school (conservatory), i.e. they complied with mandatory school attendance. Primary education is also listed by persons who completed a special school, practical school, special primary school or former junior secondary „town" school. Persons older than 15 years who at the time of the census are attending the ninth year of a primary school should also state that they have completed primary education.
Complete secondary general (with graduation) also includes lycee branches of secondary schools.
Complete secondary specialist (with graduation) also includes, apart from secondary specialist schools, vocational branches with graduation.
Further education is listed by graduates of further education or post-graduate studies, as well as graduates of two or more secondary schools completed with a final exam or graduation.
Master's includes the majority of graduates of universities; recipients of the titles Mgr, MgA, Ing., MUDr., JUDr., RNDr., MVDr., PhDr., ThDr. etc.
Doctoral - includes only current doctoral study programmes and former scientific activities, i.e. the recipients of the titles Ph.D., Th.D., CSc., DrSc.

14. FIELD OF EDUCATION
Filled out only by persons with higher than primary education. Educational branch, or vocational branch is listed according to the highest completed school. The title or specialist orientation of the graduated branch of study (e.g. biology, economics, electrotechnology, machinery, agriculture, social work, primary school teaching, public administration activity etc.) or of a vocational branch (e.g. locksmith, mechanic repairer, bricklayer, cook, salesperson, chemist-technician, textile production, food processing, timber processing, business operations, care-giving, etc.), or of a type/graduate oriented school (e.g. general education - grammar school, secondary general educational school, etc.) is entered.
If studies at more schools of an equal level have been completed, state the branch which is (or was) used in employment; if there is none, state the branch of the last completed school.

15. TOTAL NUMBER OF CHILDREN BORN ALIVE
Include in the number of children born alive up to the census day (prior to marriage, within marriage or outside it) even if some of the children later died.

16. NUMBER OF CHILDREN BORN ALIVE DURING CURRENT OR LAST MARRIAGE
State the number of children who were born up to the census day in current marriage (for married women) or in the last marriage (for divorced or widowed women), even if some of the children later died.

17. ECONOMIC ACTIVITY
The employed are all persons 15 years and older who are in paid work on census day as „employees", including „self-employed" (employers, independent operators) or assisting members of a family. The length of their work load is not important, nor whether their work activity is of a permanent, temporary, seasonal or opportune character. Independent relates to working retirees, working students and apprentice, and women on maternity leave (28 or 37 weeks).
The unemployed are all persons 15 years and older who on census day are without work, are actively looking for work and are prepared for immediate entry into work.
Economically inactive are all persons who on census day were not employed or did not meet the conditions for classification among the unemployed.
Other with own source of livelihood represents persons who live from financial sources other than income from employment or the pension (e.g. from savings from property income, dividends, social benefits etc). They include women and men on parental leave if they collect a parental benefit.
Pupils, students, apprentices represents those who dedicate themselves exclusively to studies at all types of schools.
Persons at home, pre-school age children, other dependent persons also represents adults if they are dependent for their livelihood on some member of the household.

18. OCCUPATION
Occupation represents a specific performed activity; it generally includes a work position, function and branch (e.g. car assembly worker, transport operator, mobile crane driver, furniture upholsterer, chemical production worker, service maintenance mechanic, construction electrical engineer, shop cashier, bank counter worker, pre-school teacher, business officer, garden architect, tax official, business branch manager, head nurse, primary school deputy principal, warehouse manager, hotel director, transport dispatcher, scientific worker - biochemistry, city policeman etc.). In the event of two jobs this question should be answered according to the main job.

19. STATUS IN EMPLOYMENT
Employees are persons in employment, employees appointed or elected, persons employed on the basis of a contract for work or a contract for services, employees in service (i.e. members of the army and police).
Businesspeople with employees (employers) are persons who employ one or more employees as part of their business activities.
Businesspeople without employees (self-employed persons) are persons working on their own account, and with business authorisation. They are persons entered in the business register and in the trades register (trade certificate, trade licence), persons doing business on the basis of specific regulations (e.g. advocates, experts, auditors, artists, etc.), and persons undertaking agricultural activities in accordance with specific regulations.
Members of a production co-operative are members of manufacturing, agricultural or other production co-operatives. They do not include regular employees of these co-operatives or members of consumer co-operatives.
Assisting family members are persons who work on a basis other than a labour relationship. If they are family members employed in a family enterprise on the basis of, for example, a work contract, they should be indicated as employees in employment.

20. INDUSTRY (BRANCH OF ECONOMIC ACTIVITY)
Branches are designated according to the main type of economic activity of the employer (subject of business). For example, „education" should be entered by all school employees - teachers, secretaries, and also maintenance workers, cleaners, school cafeteria employees etc., unless they are employees of a company which provides these services for the school. Please state the name of the branch of economic activity according to the following overview (or choose the relevant part of the name):

- *agriculture, hunting*
- *forestry and logging*
- *fishing and fish farming*
- *mining and preparation of coal, minerals, oil and natural gas, stone, sand, etc.; ancilliary activities for mining (research, drilling, drainage, etc.)*
- *production of foodstuff products, drinks, tobacco products*
- *production of textiles, clothing, fur products, shoes; tanning industry*
- *wood processing, manufacture of paper, pulp and products from paper, wood, wickerwork, cork, etc.*
- *print, binding and related activities; reproduction of recorded media*
- *production of coke and refined oil products*
- *production of chemical materials and medicaments*
- *pharmaceutical production*
- *production of rubber and plastic products*
- *production and processing of glass, ceramics, porcelain, stone, construction materials and products*
- *production and metallurgic processing of metals; foundry industry*
- *manufacture of metal construction and metalwork products; preparation of metals; machine tooling*
- *manufacture of computers, electronic, measuring and optic instruments and equipment, communication equipment, consumer electronics*
- *manufacture of electrical equipment (incl. distribution and control equipment, household appliances, batteries and accumulators, optic and electrical cabling, etc.)*
- *manufacture of other machinery and equipment (e.g. agricultural, cutting, for mining and industrial production, liking equipment, pumps, bearings, gearing, hand tools with motor, etc.)*
- *manufacture of motor vehicles (incl. their motors, parts and appurtenances), trailers and semi-trailers*
- *manufacture of other transport and equipment*
- *manufacture of furniture*
- *other processing industry (manufacture of jewelry, toys, musical instruments, sporting goods etc.)*
- *repairs, maintenance, and installation of machines and equipment*
- *production and distribution of electricity, gas and heating*
- *water supply*
- *activities related to waste water*
- *assembly, collection and removal of waste, waste treatment for further use*
- *sanitation and other activites related to waste*
- *building construction*
- *civil engineering (construction of roads, railways, infrastructure networks, bridges, etc)*
- *specialised construction activities (demolition, roofing, electro installation, plumbing etc, building installation work, completion and finishing work - plastering, glazing, etc.)*
- *motor vehicle trade (incl. parts and appurtenances), repairs and maintenance of motor vehicles*
- *wholesale and mediation of wholesale*
- *retail*
- *land and metro transport*
- *water transport*
- *airtransport*
- *delivery and secondary transport activities (freight and unloading of goods, operation and servicing of roads, parking, terminals - train stations, airports, transit centres, maintenance of transport facilities, etc.)*
- *postal and courier activities*
- *accommodation; food and hospitality*
- *publishing activities (publication of books, periodicals, software and so on)*
- *activities in the field of film, video recordings and TV programmes, creation of sound recordings and music publishing activities creation of television and radio programmes and broadcasting telecommunication activities*
- *activities in the field of information technology (programming, management of computer equipment, consultancy, etc.)*
- *information activities (data processing, hosting, news agency etc.) banking*
- *insurance and pension funds (other than mandatory social insurance) real estate activities (purchase, sale, renting and management of property, real estate agency activities, valuation, etc.)*
- *legal, accounting and auditing activities; tax consultancy*
- *consultancy and services for businesses in the field of management, organisation, planning, control, marketing, human resources, communication, etc.*
- *architectural and engineering activities and consultancy (project design, construction design, geological research, geodetic activities, cartography, meterology, etc.)*
- *technical tests and analysis (chem., physics etc.)*
- *research and development*
- *advertising, market research and public opinion surveys*
- *other professions, scientific and technical activities (design, graphics, photography, translation, specialist consultancy etc.)*
- *renting and leasing*
- *employment mediation*
- *activities of travel agents/offices, reservations, guide services, information and other services in the travel industry*
- *security and detective services (private security agencies, security system operations)*
- *cleaning activities*
- *landscape design (incl. gardens, parks, public areas)*

- *administrative and other ancilliary activities for business (call centres, packaging services, congress and trade fair organisation etc.)*
- *public administration (state administration, local government), defense, mandatory social insurance, public order and security, justice and courts, correctional facilities, fire brigade*
- *schools, education (incl. Education outside the school system - courses and training, driving schools, art and sports education, etc.)*
- *health care*
- *veterinary activities*
- *social welfare in facilities*
- *clinic and field social services*
- *creative, artistic and entertainment activities (artistic design, theatre, musical etc., art, cultural facility operations)*
- *activities of libraries, archives, museums, botanic and zoological gardens, natural reserves, service and protection of heritage sites*
- *gambling, casino and betting shop activities*
- *sports, entertainment and recreational activities*
- *activities of professional, political, social, cultural, interest organisations, churches and religious organisations, trade unions, movements and other associations*
- *repairs of computers and products for personal requirements and primarily for the household*
- *other personal services (dry cleaning, haidressing, beautician etc.)*
- *household personnel, individual (child care, household help, personal driver, butler etc.)*
- *international organisation activities (UN, EU, World Bank etc.)*

Questions 21,22,23 and 24 on commuting to work or school are filled out by persons who stated in the question on economic activity that they are employed or are pupils, students and apprentices (within the scope of economically inactive persons).
„Working students and apprentices" concerns information about commuting to school, not information about commuting to work. Commuting is also considered to be commuting within a town or city.

21. LOCATION OF PLACE OF WORK OR SCHOOL
State the address of the place of work (not, for example, the company headquarters).
Persons who have no workplace but start work at the same address (e.g. transport workers - drivers, pilots; artisans-repairers etc.), state where they start work.
Persons who often change their place of work (e.g. assembly and construction workers, marketplace sales people etc.), or travel but do not come to the same address daily (travelling businesspeople, taxi drivers, truck drivers etc.) should indicate „employed without permanent workplace".
Persons working or studying abroad should state the name of the state.
Persons whose workplace address (schools) is the same as the address of their (actual) residence, and employees without a permanent workplace do not fill out any further questions on commuting.

Questions 22 and 23 relate to daily journey to work or school (from home or from a place of temporary accommodation).
Persons working in shifts enter information regarding a single journey to work even if they do not commute regularly every day.

22. MODE OF TRANSPORT
Bus (other than city transport) is stated by persons using bus transport which crosses the border of a municipality/city. It includes suburban transport.
Mass city transport is mass public transport operated for the satisfaction of the transport requirements for a city area. In Prague, it includes the metro.
Other means of transport includes all others, namely those not listed on the form.
None (on foot only) is stated by persons who walk to work or school, i.e. they do not use any means of transport.

23. TIME SPENT ON DAILY JOURNEY
State the duration of one normal journey to work or school. The time in minutes includes the total time which elapses from leaving home (or a place of temporary accommodation) to entry to the workplace or entry to school („door to door"), i.e. including the walk to a public transport station and from a station, waiting for arrival, change etc.

24. FREQUENCY OF JOURNEY TO WORK OR SCHOOL
States how often a person commutes from their municipality of residence to the municipality of their workplace or school. Persons whose workplace or school is in the same town of (actual) residence do not answer the question.
Daily is stated by persons who commute directly from their place of (actual) residence and do not use temporary accommodation at the site of the workplace or school (rental, boarding house, student hostel, school hostel, etc.). They may also be persons working in shifts or for a shortened time if they meet the stated conditions, even if they do not commute regularly every day.

EXPLANATIONS FOR THE CENSUS FORM - DWELLINGS are on the other side. Please turn over.

EXPLANATIONS FOR THE CENSUS FORM - BUILDINGS are on the back page of the Form. The Census Form – Buildings is filled out by building owners or managers.

Any ambiguities and questions on filling out the census forms will be answered by the census officer, or he/she will help you fill out the forms. He/she will provide you with all other required information. You can also get further information at the internet page www.scitani.cz, and from the free-of-charge census telephone information line 800 879 702 or by emailing info@scitani.cz.

[B] Correlation matrices of explanatory variables

B.1 Municipality-related variables

Correlation matrix of municipality-related explanatory variables	x100	x110	x120	x130	x140	x200	x210	x220	x230	x240	x250	x260	x270	x280	x300	x310	x320	x330	x340	x350	x360	x370	x400	x410	x420	x430	x500	x510	x600	x610	x700	x710	x720	x730	x800	x810	x820	x905	x915
Average age of inhabitants - **x100**																																							
Share of children aged 0-14 - **x110**	-0,77																																						
Share of inhabitants aged 15-64 - **x120**	-0,33	-0,24																																					
Share of inhabitants with age 65+ - **x130**	0,86	-0,51	-0,70																																				
Age index (share of 65+ / share of 0-14) - **x140**	0,75	-0,66	-0,25	0,70																																			
Share of employees with elementary education - **x200**	-0,06	-0,01	0,16	-0,14	-0,07																																		
Share of employees with apprenticeship certificate - **x210**	0,15	-0,12	-0,07	0,15	0,10	0,14																																	
Share of employees without high school diploma - **x220**	0,10	-0,11	0,01	0,07	0,06	0,53	0,91																																
Share of employees with high school diploma - **x230**	-0,05	0,04	-0,01	-0,02	-0,03	-0,43	-0,69	-0,77																															
Share of employees with completed extension study - **x240**	-0,01	0,03	-0,04	0,01	-0,01	-0,19	-0,27	-0,31	0,08																														
Share of employees with tertiary technical education - **x250**	-0,03	0,03	-0,02	0,00	-0,01	-0,15	-0,23	-0,26	0,07	0,09																													
Share of employees with university degree - **x260**	-0,11	0,12	0,01	-0,09	-0,06	-0,38	-0,71	-0,76	0,24	0,14	0,14																												
Share of employees with high school diploma or higher - **x270**	-0,10	0,11	-0,01	-0,06	-0,05	-0,54	-0,90	-0,99	0,77	0,31	0,26	0,77																											
Average length of the education attendance of employees - **x280**	-0,08	0,11	-0,04	-0,04	-0,03	-0,64	-0,73	-0,88	0,44	0,28	0,24	0,94	0,89																										
Unemployment rate - **x300**	-0,01	-0,06	0,13	-0,07	0,00	0,30	0,22	0,31	-0,24	-0,09	-0,12	-0,24	-0,32	-0,32																									
Share of economically active inhabitants - **x310**	-0,30	-0,02	0,54	-0,45	-0,22	-0,07	-0,20	-0,20	0,16	0,05	0,06	0,16	0,21	0,17	-0,02																								
Share of retired - **x320**	0,86	-0,55	-0,48	0,84	0,64	-0,05	0,26	0,21	-0,13	-0,04	-0,04	-0,20	-0,21	-0,17	0,03	-0,58																							
Share of pupils and students - **x330**	-0,50	0,48	-0,03	-0,32	-0,39	-0,18	-0,02	-0,09	0,06	0,02	-0,01	0,09	0,09	0,13	-0,04	0,01	-0,38																						
Ratio of economically inactive to active inhabitants - **x340**	0,30	-0,03	-0,47	0,43	0,25	0,06	0,17	0,17	-0,14	-0,04	-0,02	-0,14	-0,17	-0,15	-0,01	-0,94	0,54	-0,07																					
Ratio of self-employed to economically active inhabitants - **x350**	0,03	0,05	-0,10	0,05	0,02	-0,13	-0,15	-0,18	0,12	0,01	0,06	0,17	0,18	0,19	-0,16	0,02	-0,03	-0,03	-0,01																				
Ratio of occupied jobs to economically active inhabitants - **x360**	-0,04	-0,02	0,08	-0,07	-0,04	0,07	-0,12	-0,07	0,03	0,05	0,01	0,07	0,06	0,04	-0,09	-0,01	-0,08	-0,04	0,03	0,00																			
Ratio of inhabitants employed in key sectors (PS-/T+) - **x370**	-0,17	0,13	0,11	-0,19	-0,12	-0,10	-0,54	-0,51	0,29	0,14	0,14	0,49	0,50	0,47	-0,12	0,17	-0,28	-0,06	-0,13	0,22	0,12																		
Share of municipal inhabitant on national tax revenue - **x400**	0,24	-0,17	-0,04	0,16	0,20	0,14	0,11	0,15	-0,12	-0,05	-0,05	-0,11	-0,15	-0,15	0,16	-0,10	0,20	-0,17	0,09	0,13	0,01	-0,08																	
Car ownership - **x410**	0,16	-0,14	0,05	0,06	0,14	-0,04	0,03	0,01	0,03	-0,02	0,03	-0,04	-0,01	-0,02	-0,14	0,20	0,03	-0,21	-0,19	0,18	0,03	-0,02	0,25																
Number of inhabitants - **x420**	0,00	-0,02	0,02	0,00	0,00	-0,02	-0,11	-0,10	0,04	0,03	0,02	0,11	0,10	0,10	-0,01	0,02	-0,03	-0,02	-0,03	-0,01	0,08	0,12	0,19	-0,05															
Total number of workers, pupils and students - **x430**	0,00	-0,02	0,02	0,00	-0,01	-0,02	-0,11	-0,10	0,04	0,03	0,03	0,11	0,10	0,10	-0,01	0,03	-0,03	-0,02	-0,03	-0,01	0,08	0,12	0,19	-0,05	1,00														
Gross population density - **x500**	-0,12	0,05	0,07	-0,10	-0,09	-0,08	-0,35	-0,33	0,18	0,11	0,07	0,33	0,33	0,32	-0,07	0,12	-0,17	0,02	-0,12	-0,07	0,17	0,28	-0,10	-0,23	0,44	0,44													
Net population density - **x510**	-0,27	0,16	0,12	-0,23	-0,20	-0,10	-0,36	-0,35	0,21	0,12	0,06	0,34	0,35	0,34	-0,05	0,17	-0,28	0,16	-0,17	-0,11	0,09	0,27	-0,24	-0,41	0,26	0,26	0,74												
Municipality organizing Urban Public Transport - **x600**	0,01	-0,04	0,03	0,00	-0,01	-0,01	-0,18	-0,16	0,10	0,07	0,04	0,13	0,15	0,13	0,00	0,04	-0,04	-0,03	-0,04	-0,05	0,16	0,15	0,07	-0,11	0,32	0,31	0,54	0,41											
Distance to the closest train stop or station - **x610**	0,17	-0,13	-0,09	0,17	0,17	0,01	0,19	0,16	-0,15	-0,06	-0,04	-0,09	-0,16	-0,10	0,06	-0,10	0,19	-0,04	0,10	0,08	-0,08	-0,11	0,23	0,12	-0,06	-0,06	-0,25	-0,29	-0,12										
Average purchase price of house - **x700**	-0,19	0,16	0,07	-0,18	-0,12	-0,07	-0,41	-0,38	0,19	0,06	0,11	0,41	0,38	0,38	-0,24	0,21	-0,30	-0,09	-0,18	0,23	0,09	0,56	-0,14	0,10	0,14	0,14	0,32	0,28	0,13	-0,13									
Average purchase price of apartment - **x710**	-0,11	0,10	-0,01	-0,06	-0,07	-0,20	-0,35	-0,38	0,19	0,09	0,13	0,40	0,38	0,41	-0,27	0,18	-0,19	-0,04	-0,16	0,18	0,07	0,41	-0,18	-0,01	0,15	0,16	0,31	0,26	0,13	-0,09	0,81								
Average purchase price of building land - **x720**	-0,15	0,12	0,05	-0,13	-0,10	-0,08	-0,41	-0,39	0,18	0,08	0,11	0,43	0,39	0,39	-0,17	0,16	-0,25	-0,05	-0,14	0,14	0,15	0,52	-0,07	-0,05	0,37	0,37	0,53	0,44	0,37	-0,15	0,85	0,75							
Share of unoccupied dwellings - **x730**	0,43	-0,25	-0,24	0,39	0,34	0,02	0,21	0,19	-0,14	-0,04	0,01	-0,18	-0,19	-0,16	0,03	-0,28	0,43	-0,23	0,29	0,20	-0,10	-0,15	0,36	0,33	-0,09	-0,09	-0,41	-0,59	-0,17	0,32	-0,17	-0,17	-0,24						
Mean altitude of municipality - **x800**	0,11	-0,07	-0,04	0,09	0,09	0,05	0,09	0,10	-0,08	0,00	0,01	-0,08	-0,10	-0,09	0,04	-0,12	0,13	0,03	0,11	0,08	0,02	-0,13	0,35	0,02	-0,02	-0,02	-0,19	-0,06	-0,01	0,21	-0,23	-0,28	-0,17	0,29					
Mountainousness of municipality - **x810**	0,03	-0,04	0,06	-0,03	-0,03	0,08	-0,06	-0,02	0,00	0,04	-0,04	0,02	0,01	0,00	0,14	-0,05	0,02	0,00	0,03	0,04	0,07	0,06	0,20	-0,15	0,05	0,04	-0,01	0,18	0,09	0,05	-0,08	-0,12	-0,01	0,06	0,62				
Mountainousness of municipality relative to its area - **x820**	0,09	-0,04	-0,08	0,10	0,09	-0,12	-0,04	-0,08	0,05	0,04	0,02	0,07	0,08	0,10	-0,03	-0,06	0,09	0,01	0,07	0,12	-0,10	0,01	-0,08	0,03	-0,06	-0,06	-0,02	0,03	-0,11	0,04	-0,04	-0,04	-0,07	0,14	0,22	0,25			
Annual average precipitation (1991-2010) - **x905**	0,06	-0,03	-0,04	0,05	0,03	-0,03	0,06	0,03	-0,02	0,03	0,01	-0,05	-0,04	-0,03	0,00	-0,10	0,08	0,11	0,08	0,02	0,03	-0,18	0,18	-0,10	-0,02	-0,02	-0,08	0,07	0,00	0,06	-0,24	-0,25	-0,18	0,18	0,64	0,50	0,22		
Annual average temperature (1991-2010) - **x915**	-0,09	0,06	0,03	-0,06	-0,06	-0,05	-0,09	-0,10	0,08	0,00	-0,01	0,10	0,10	0,10	-0,06	0,13	-0,11	-0,05	-0,12	-0,05	-0,03	0,14	-0,30	0,02	0,05	0,05	0,18	0,04	0,04	-0,15	0,24	0,30	0,18	-0,26	-0,88	-0,54	-0,23	-0,67	

Note: The ones on the diagonal have been deleted not to distort the colour scale of conditional formatting.

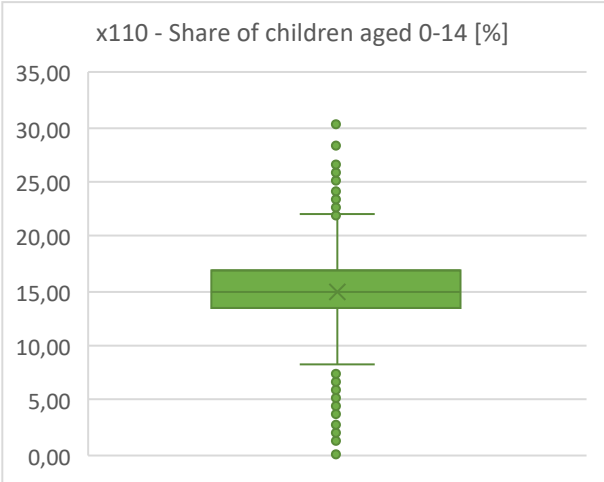
B.2 Variables related to connections of municipalities

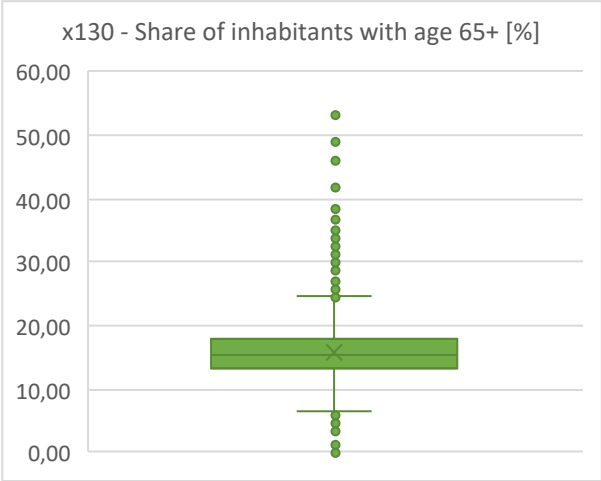
Variables related to connections of municipalities	x10	x11	x12	x13	x20	x21	x22	x23	x31	x32	x34	x40	x41	x42	x43	x44	x45	x60	x61	x62	x63	x70	x71	x73	x74	x76	x80	x81	x83	x84	x86	x90	x91
Distance to Bus stop - **x10**																																	
Distance from Bus stop - **x11**	-0.11																																
Total distance to and from Bus stops - **x12**	0.65	0.68																															
Number of connections by Bus - **x13**	-0.06	0.08	0.02																														
Distance to Train stop or station - **x20**	-0.12	0.06	-0.04	0.03																													
Distance from Train stop or station - **x21**	0.00	-0.24	-0.18	-0.05	-0.10																												
Total distance to and from Train stops or stations - **x22**	-0.08	-0.13	-0.16	-0.02	0.65	0.69																											
Number of connections by Train - **x23**	0.15	0.14	0.22	0.12	-0.26	-0.24	-0.38																										
Distance by I. class road - **x31**	-0.05	0.00	-0.04	0.13	0.06	-0.09	-0.02	-0.01																									
Distance by II. or III. class road - **x32**	-0.06	0.01	-0.04	-0.31	0.14	0.00	0.10	0.07	-0.02																								
Proportion of distance by I. class road - **x34**	-0.03	0.02	-0.01	0.37	0.04	-0.01	0.02	0.00	0.75	-0.35																							
Shortcut / prolonging by Bike - **x40**	-0.08	0.04	-0.03	-0.16	-0.04	-0.12	-0.12	-0.04	0.23	0.41	0.09																						
Shortcut / prolonging by Bike relative - **x41**	-0.08	0.04	-0.03	-0.06	-0.10	-0.16	-0.20	-0.10	0.08	0.07	0.10	0.73																					
Distance by rural cycle path - **x42**	0.09	0.06	0.12	0.38	-0.11	-0.06	-0.12	0.27	0.12	-0.07	0.28	-0.10	-0.14																				
Distance by rural cycle path relative - **x43**	0.11	0.06	0.12	0.33	-0.12	-0.10	-0.16	0.06	-0.02	-0.18	0.10	-0.13	-0.15	0.67																			
Total elevation of Bike route - **x44**	-0.07	-0.02	-0.07	-0.28	0.24	0.06	0.22	-0.24	0.55	0.61	0.21	0.41	0.14	-0.19	-0.24																		
Total elevation of Bike route relative - **x45**	-0.01	0.05	0.03	-0.25	0.19	0.16	0.26	-0.47	0.14	-0.01	0.22	0.21	0.29	-0.32	-0.27	0.54																	
Air distance - **x60**	-0.05	0.01	-0.03	-0.14	0.16	-0.06	0.07	0.08	0.61	0.75	0.19	0.29	-0.05	0.05	-0.14	0.77	0.01																
Total distance by roads - **x61**	-0.08	0.00	-0.05	-0.18	0.15	-0.05	0.07	0.05	0.58	0.80	0.16	0.47	0.10	0.02	-0.16	0.82	0.08	0.97															
Total distance by Bike - **x62**	-0.07	0.00	-0.05	-0.16	0.17	-0.04	0.10	0.06	0.58	0.78	0.16	0.32	-0.03	0.04	-0.15	0.80	0.04	0.99	0.99														
Level of neighbouring - **x63**	-0.03	0.09	0.04	-0.15	0.07	-0.01	0.04	0.35	0.23	0.61	-0.07	0.07	-0.13	0.06	-0.15	0.32	-0.22	0.69	0.64	0.67													
Travel time by Bus - **x70**	-0.03	0.24	0.16	0.17	0.20	-0.11	0.07	0.03	0.07	-0.16	0.17	-0.11	-0.09	0.00	0.05	0.06	0.20	-0.07	-0.09	-0.07	0.07												
Travel time by Train - **x71**	0.03	0.13	0.12	0.17	-0.27	-0.19	-0.34	0.17	0.00	0.05	0.00	-0.05	-0.07	0.18	0.11	-0.09	-0.16	0.07	0.04	0.05	0.10	0.10											
Travel time by Driver - **x73**	-0.12	0.21	0.07	0.02	0.06	-0.01	0.04	0.06	0.16	0.34	0.09	0.12	0.01	0.15	-0.05	0.28	0.07	0.37	0.37	0.38	0.37	0.01	0.03										
Travel time by CarPass - **x74**	0.12	0.25	0.28	0.01	0.10	-0.10	-0.01	0.15	-0.01	-0.03	0.08	0.08	0.06	0.04	-0.04	-0.05	0.05	-0.03	-0.03	-0.05	0.07	0.10	0.03	0.07									
Travel time by Bike - **x76**	0.02	0.07	0.07	0.02	-0.16	-0.16	-0.24	0.14	-0.14	-0.17	-0.14	-0.12	-0.04	0.01	0.16	-0.27	-0.28	-0.19	-0.22	-0.21	-0.08	-0.03	0.15	-0.12	0.03								
Travel time share of Bus - **x80**	-0.07	0.19	0.10	0.20	0.17	0.01	0.13	-0.09	0.04	-0.29	0.21	-0.14	-0.06	-0.01	0.17	-0.04	0.26	-0.20	-0.22	-0.21	-0.08	0.76	-0.04	-0.11	0.16	0.04							
Travel time share of Train - **x81**	-0.02	0.09	0.06	0.08	-0.33	-0.22	-0.41	0.20	-0.04	0.01	-0.03	-0.02	0.00	0.10	0.07	-0.12	-0.15	0.00	-0.02	-0.01	-0.01	-0.10	0.79	-0.01	0.05	0.22	-0.13						
Travel time share of Driver - **x83**	-0.10	0.17	0.05	0.18	0.08	-0.04	0.02	0.09	-0.07	-0.03	0.02	-0.07	-0.07	0.12	0.08	-0.12	-0.04	-0.06	-0.07	-0.06	0.01	0.02	-0.04	0.58	0.01	0.02	-0.02	-0.13					
Travel time share of CarPass - **x84**	0.12	0.29	0.31	0.07	0.03	-0.13	-0.08	0.14	-0.08	-0.14	0.06	-0.01	0.05	0.04	-0.01	-0.15	0.06	-0.15	-0.17	-0.18	-0.07	0.10	0.09	-0.01	0.77	0.14	0.25	0.15	0.09				
Travel time share of Bike - **x86**	-0.02	-0.06	-0.06	0.12	-0.12	-0.06	-0.14	-0.01	-0.11	-0.20	-0.12	-0.14	-0.11	0.04	0.21	-0.26	-0.29	-0.20	-0.23	-0.21	-0.15	-0.01	0.08	-0.22	-0.03	0.70	0.04	0.12	-0.12	0.07			
Averaged annual average precipitation - **x90**	-0.02	0.09	0.06	-0.17	0.31	0.18	0.37	-0.33	0.19	0.14	0.22	0.19	0.06	-0.20	-0.13	0.53	0.67	0.19	0.23	0.21	-0.08	0.24	-0.07	0.17	0.08	-0.23	0.21	-0.06	0.06	0.02	-0.28		
Averaged annual average temperature - **x91**	0.10	0.01	0.08	0.18	-0.22	-0.12	-0.25	0.33	-0.28	-0.15	-0.32	-0.21	-0.10	0.20	0.15	-0.57	-0.68	-0.26	-0.28	-0.27	0.08	-0.15	0.02	-0.19	-0.02	0.21	-0.14	-0.03	-0.02	0.04	0.29	-0.91	

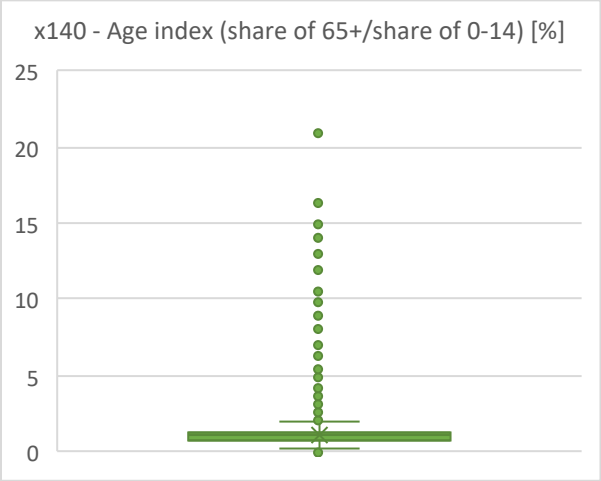
Note: The ones on the diagonal have been deleted not to distort the colour scale of conditional formatting.

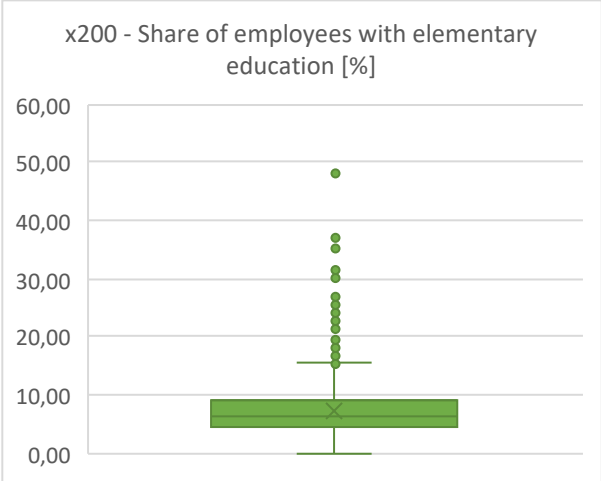
[C] Box and Whisker plots of explanatory variables

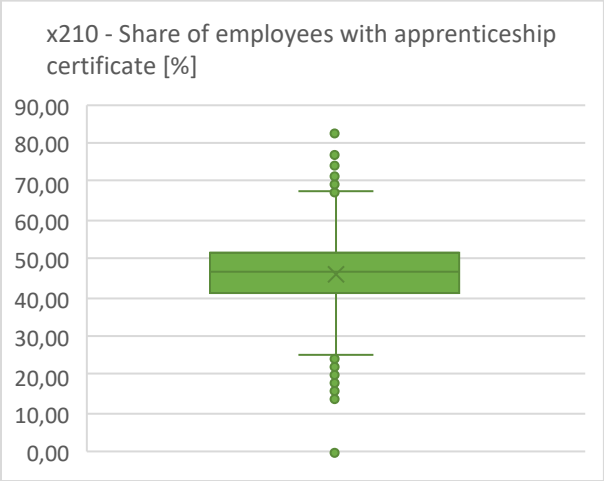
C.1 Municipality-related variables

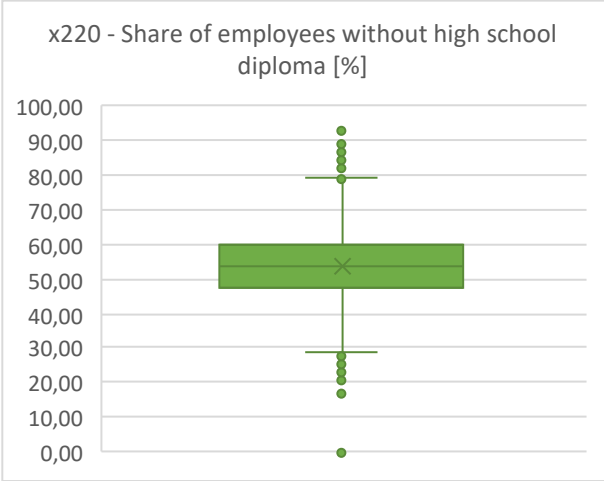


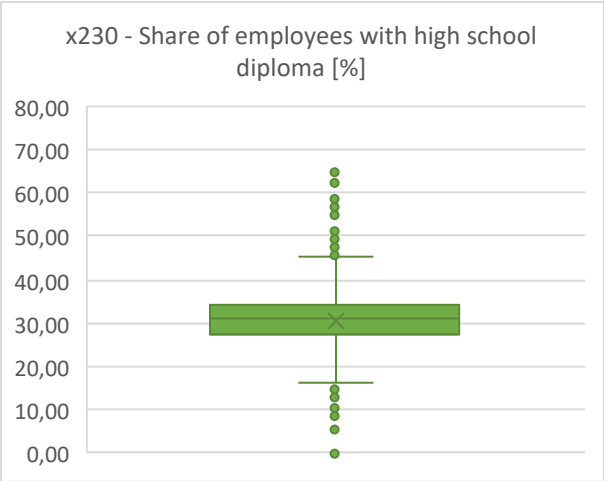


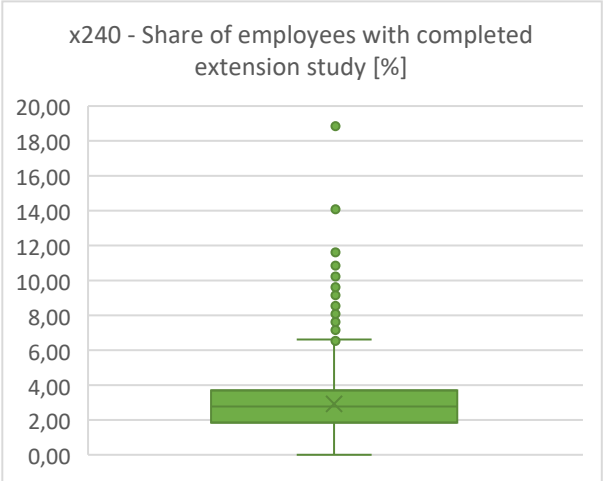


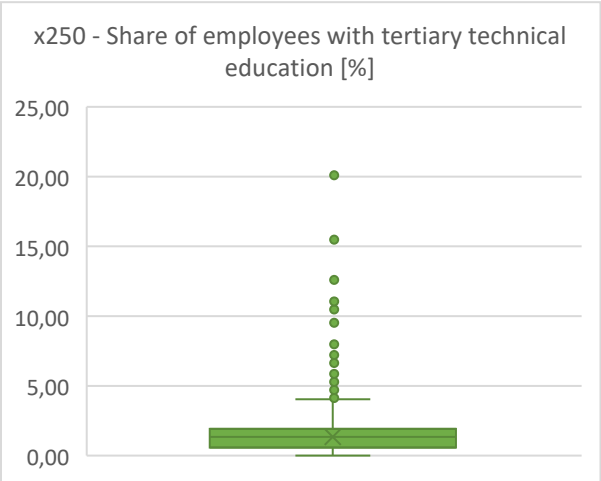


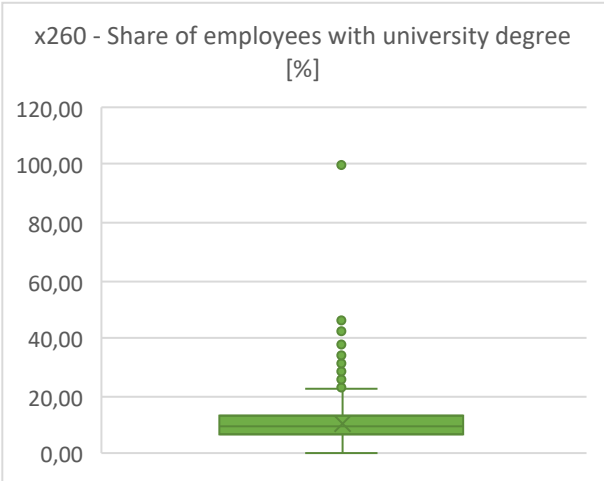

x210 - Share of employees with apprenticeship certificate [%]
90,00
80,00
70,00
60,00
50,00
40,00
30,00
20,00
10,00
0,00

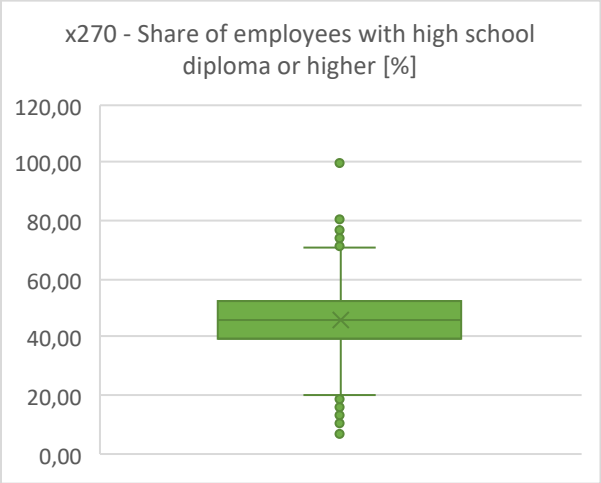

x220 - Share of employees without high school diploma [%]
100,00
90,00
80,00
70,00
60,00
50,00
40,00
30,00
20,00
10,00
0,00

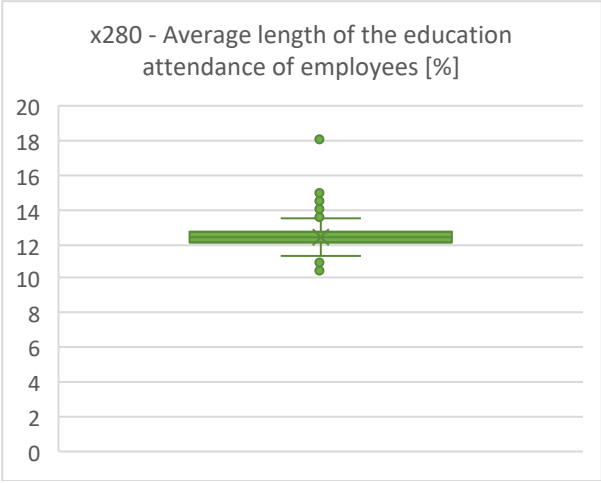

x230 - Share of employees with high school diploma [%]
80,00
70,00
60,00
50,00
40,00
30,00
20,00
10,00
0,00

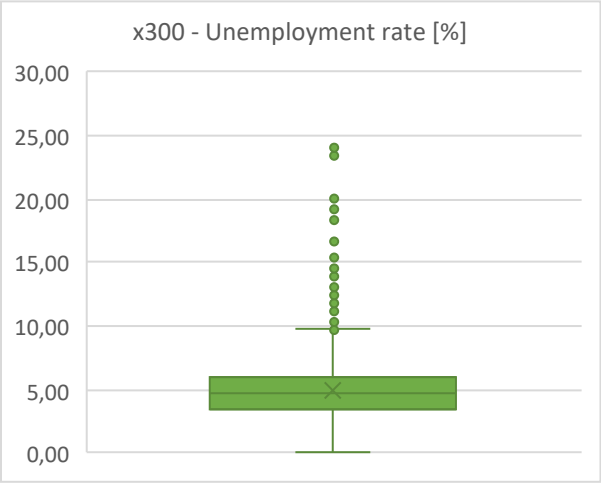

x240 - Share of employees with completed extension study [%]
20,00
18,00
16,00
14,00
12,00
10,00
8,00
6,00
4,00
2,00
0,00

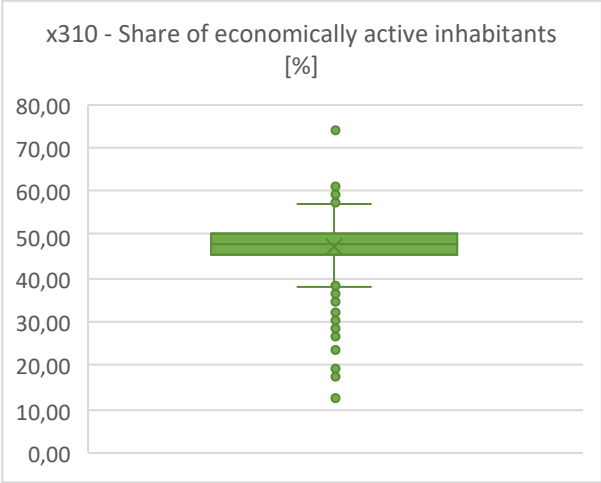

x250 - Share of employees with tertiary technical education [%]
25,00
20,00
15,00
10,00
5,00
0,00

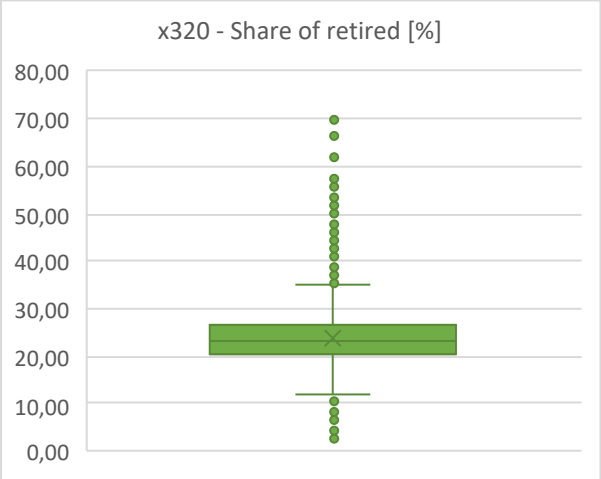

x260 - Share of employees with university degree [%]
120,00
100,00
80,00
60,00
40,00
20,00
0,00

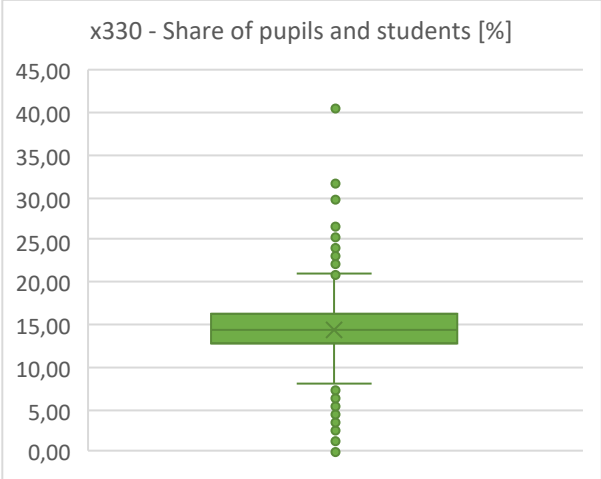

x270 - Share of employees with high school diploma or higher [%]
120,00
100,00
80,00
60,00
40,00
20,00
0,00

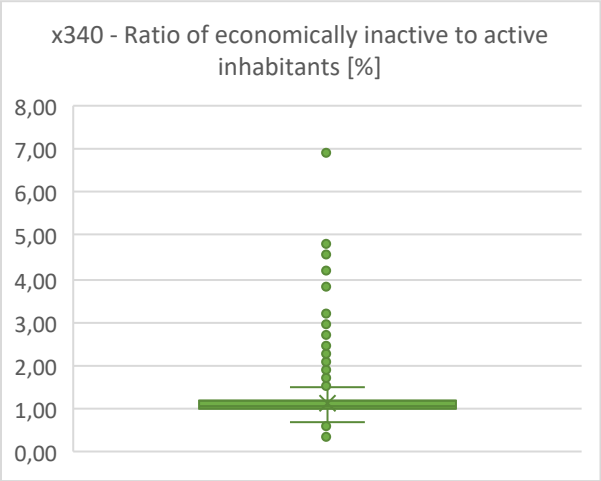

x280 - Average length of the education attendance of employees [%]
20
18
16
14
12
10
8
6
4
2
0

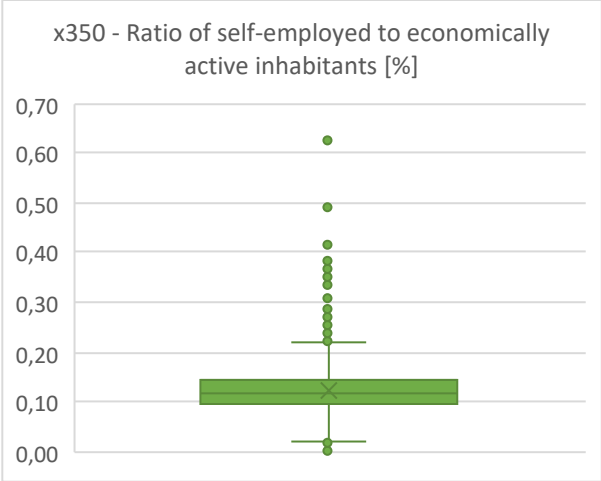

x300 - Unemployment rate [%]
30,00
25,00
20,00
15,00
10,00
5,00
0,00

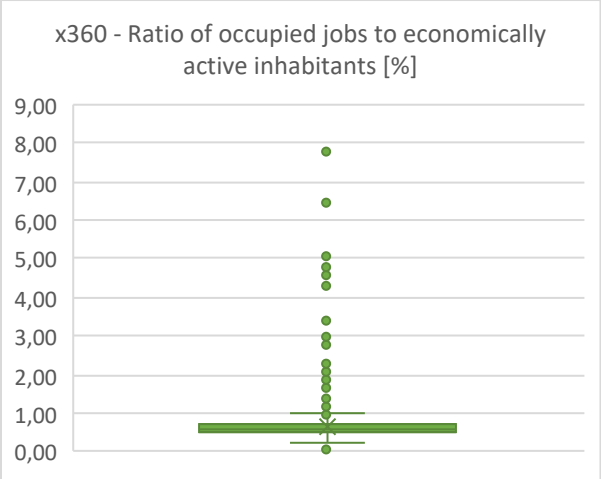

x310 - Share of economically active inhabitants [%]
80,00
70,00
60,00
50,00
40,00
30,00
20,00
10,00
0,00

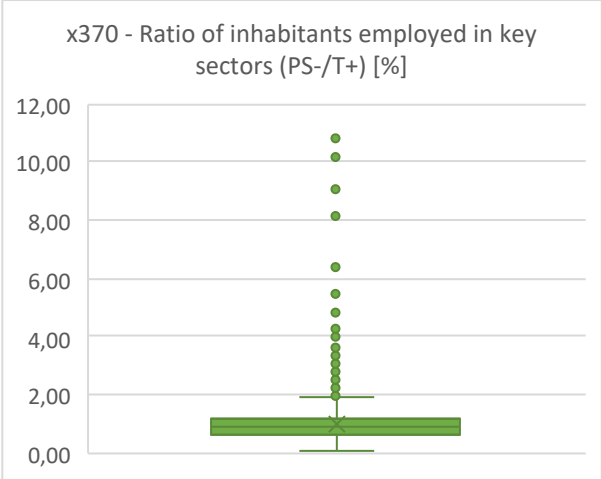

x320 - Share of retired [%]
80,00
70,00
60,00
50,00
40,00
30,00
20,00
10,00
0,00

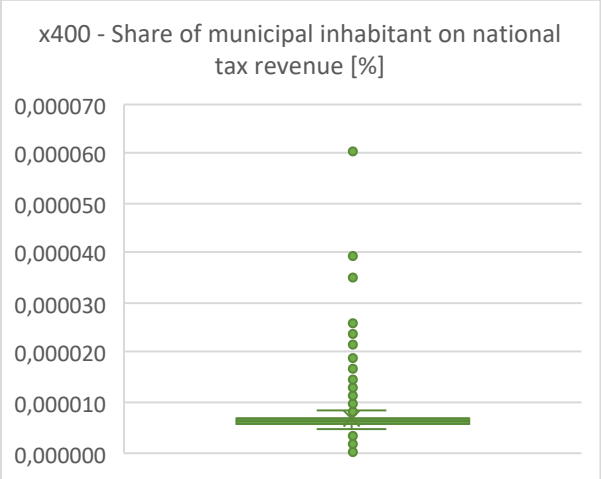

x330 - Share of pupils and students [%]
45,00
40,00
35,00
30,00
25,00
20,00
15,00
10,00
5,00
0,00

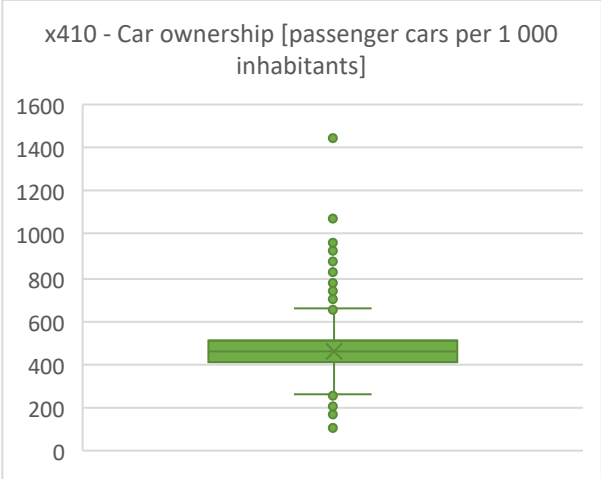

x340 - Ratio of economically inactive to active inhabitants [%]
8,00
7,00
6,00
5,00
4,00
3,00
2,00
1,00
0,00

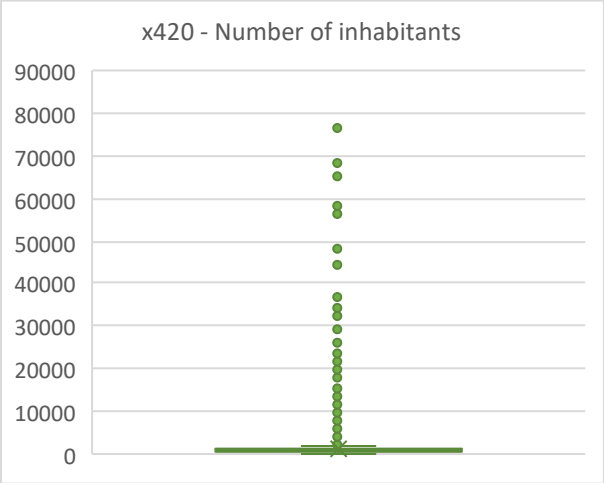

x350 - Ratio of self-employed to economically active inhabitants [%]
0,70
0,60
0,50
0,40
0,30
0,20
0,10
0,00

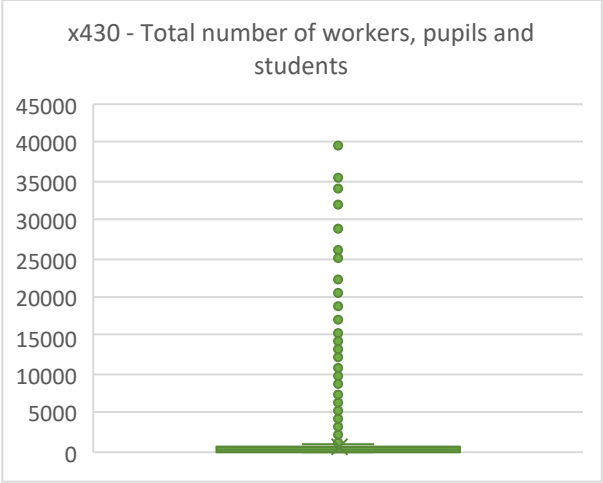

x360 - Ratio of occupied jobs to economically active inhabitants [%]
9,00
8,00
7,00
6,00
5,00
4,00
3,00
2,00
1,00
0,00

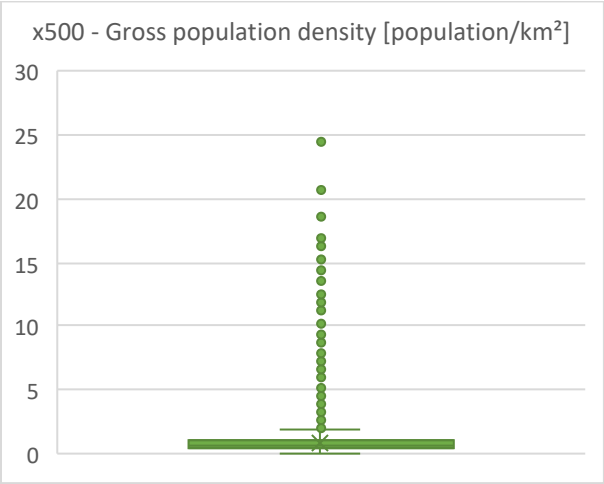

x370 - Ratio of inhabitants employed in key sectors (PS-/T+) [%]
12,00
10,00
8,00
6,00
4,00
2,00
0,00

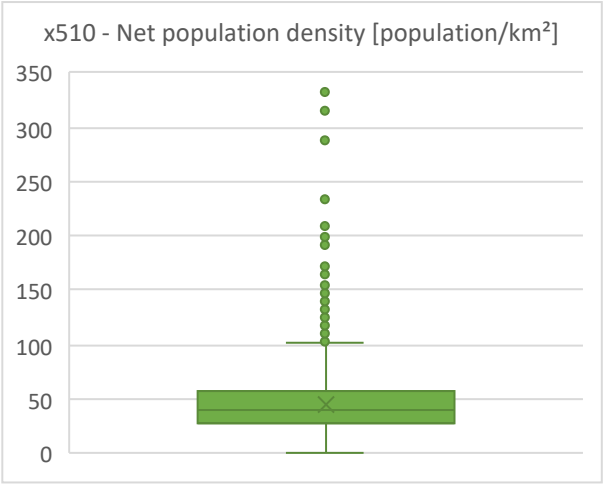

x400 - Share of municipal inhabitant on national tax revenue [%]
0,000070
0,000060
0,000050
0,000040
0,000030
0,000020
0,000010
0,000000

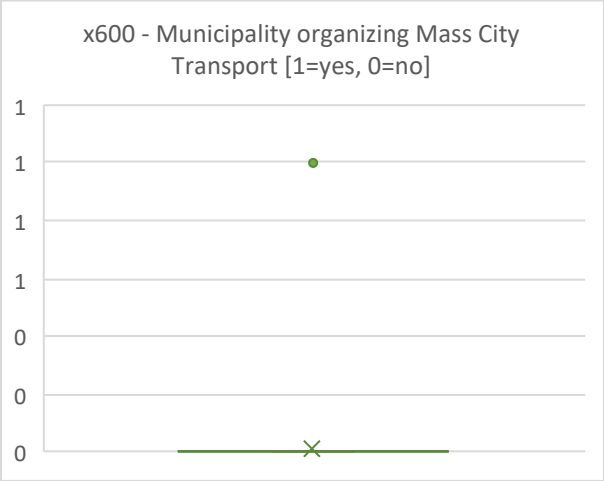

x410 - Car ownership [passenger cars per 1 000 inhabitants]
1600
1400
1200
1000
800
600
400
200
0

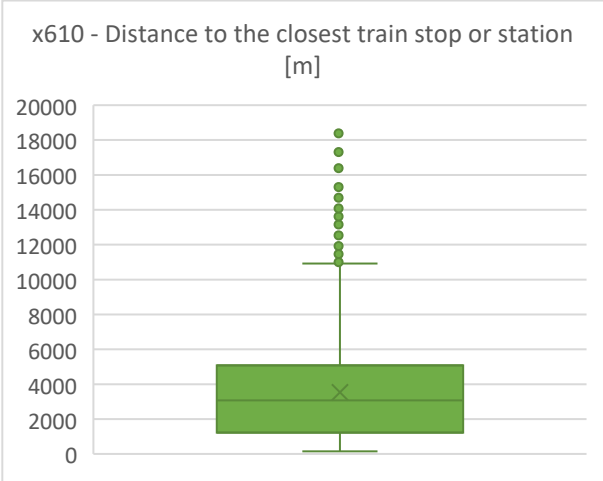

x420 - Number of inhabitants
90000
80000
70000
60000
50000
40000
30000
20000
10000
0

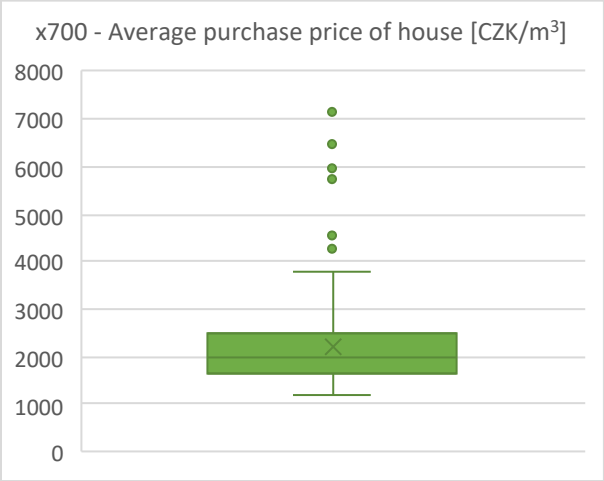

x430 - Total number of workers, pupils and students
45000
40000
35000
30000
25000
20000
15000
10000
5000
0

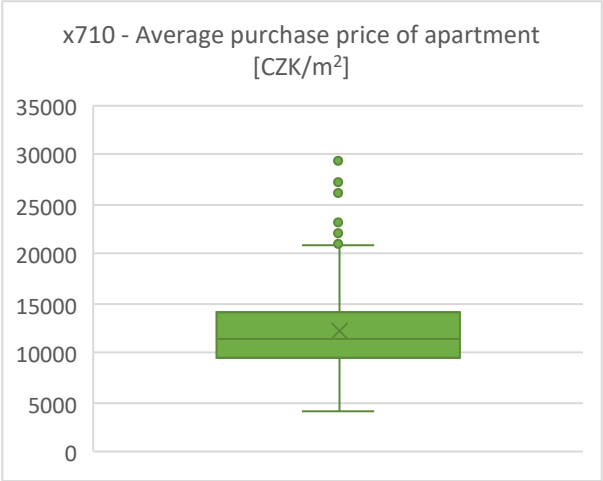

x500 - Gross population density [population/km²]
30
25
20
15
10
5
0

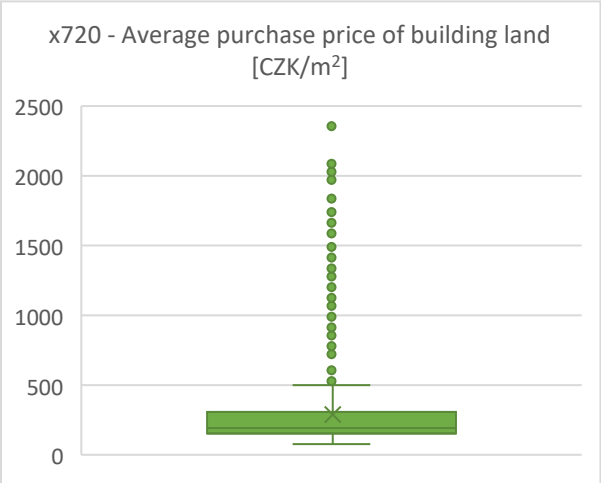

x510 - Net population density [population/km²]
350
300
250
200
150
100
50
0

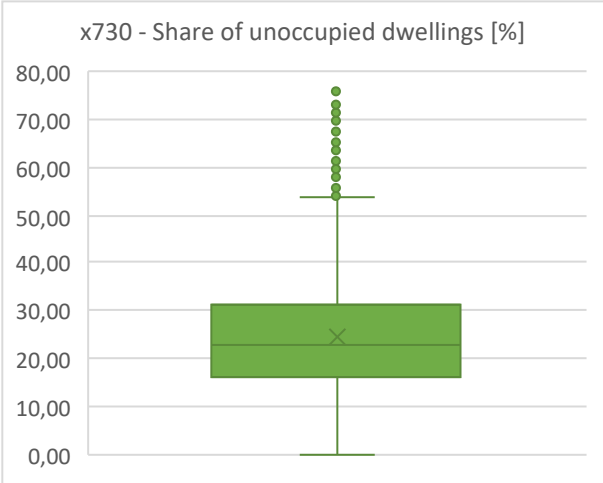

x600 - Municipality organizing Mass City Transport [1=yes, 0=no]
1
1
1
1
0
0
0

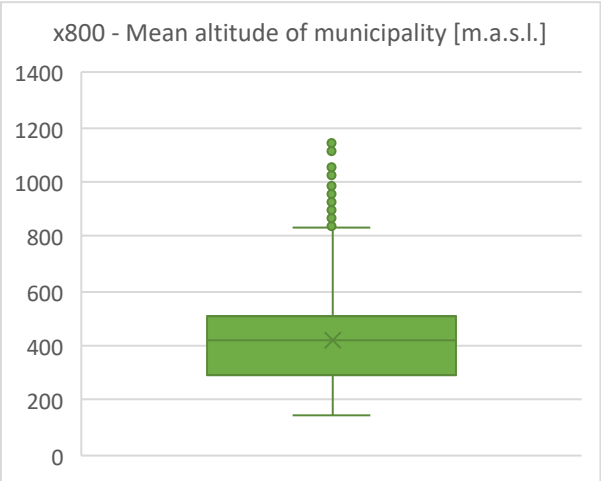

x610 - Distance to the closest train stop or station [m]
20000
18000
16000
14000
12000
10000
8000
6000
4000
2000
0

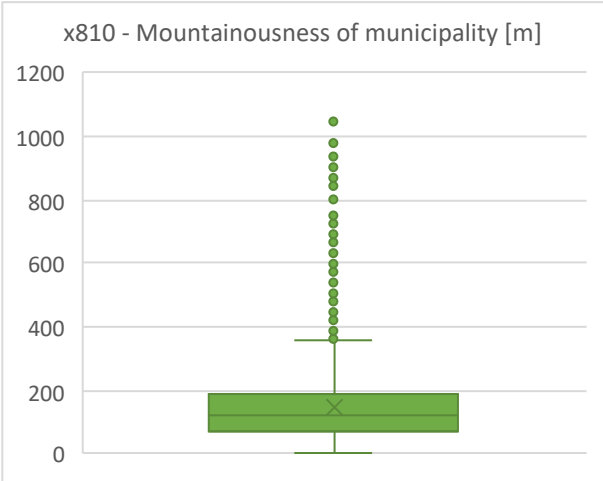

x700 - Average purchase price of house [CZK/m3]
8000
7000
6000
5000
4000
3000
2000
1000
0

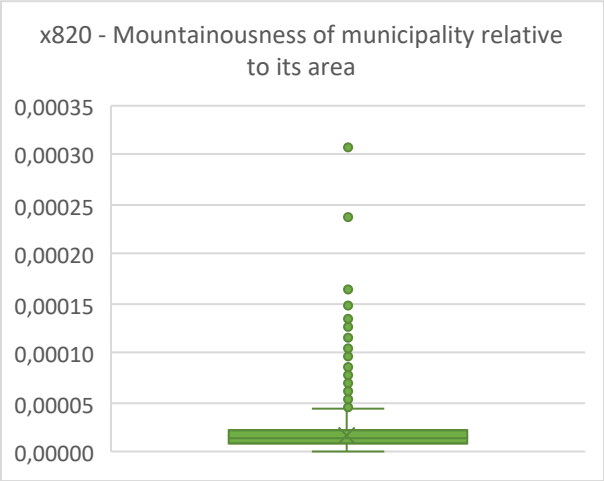

x710 - Average purchase price of apartment [CZK/m2]
35000
30000
25000
20000
15000
10000
5000
0

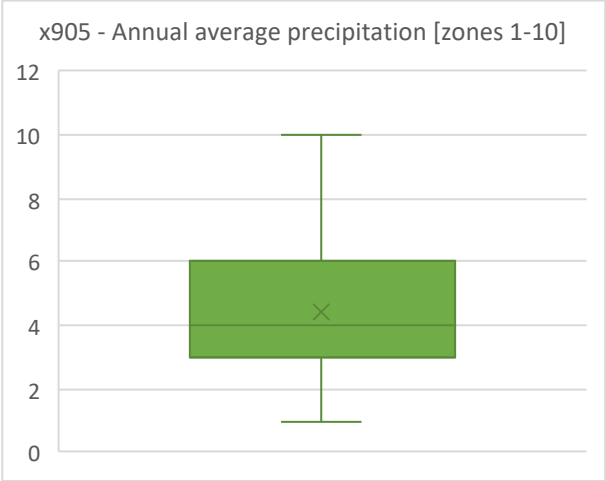

x720 - Average purchase price of building land [CZK/m2]
2500
2000
1500
1000
500
0

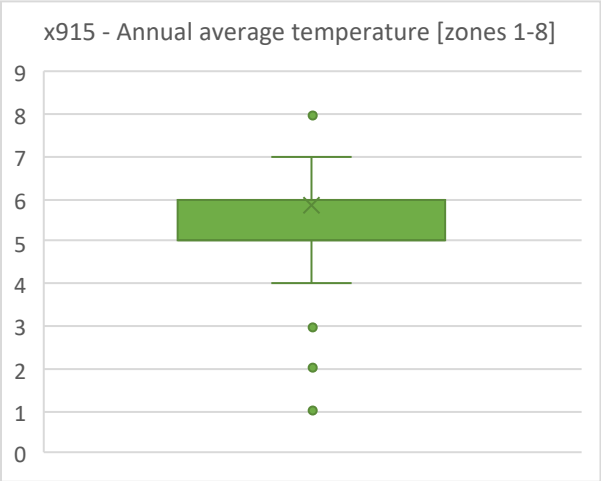

x730 - Share of unoccupied dwellings [%]
80,00
70,00
60,00
50,00
40,00
30,00
20,00
10,00
0,00


x800 - Mean altitude of municipality [m.a.s.l.]
1400
1200
1000
800
600
400
200
0

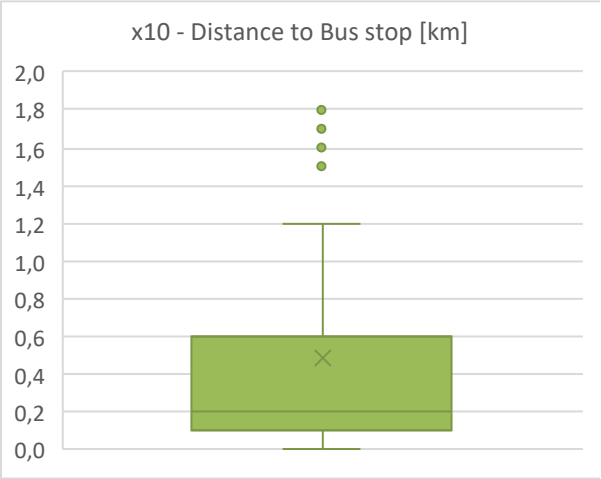

x810 - Mountainousness of municipality [m]
1200
1000
800
600
400
200
0

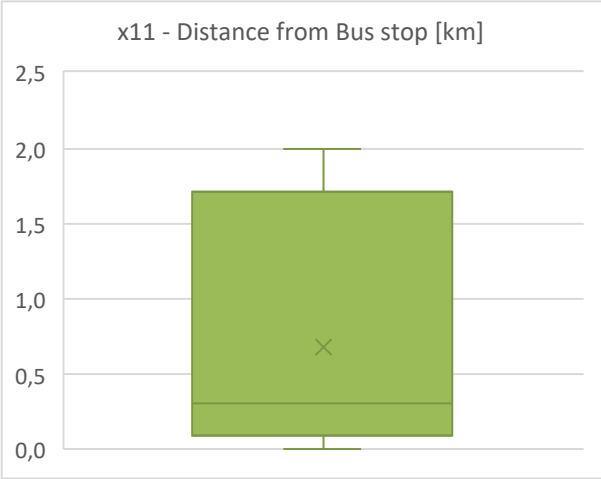

x820 - Mountainousness of municipality relative to its area
0,00035
0,00030
0,00025
0,00020
0,00015
0,00010
0,00005
0,00000

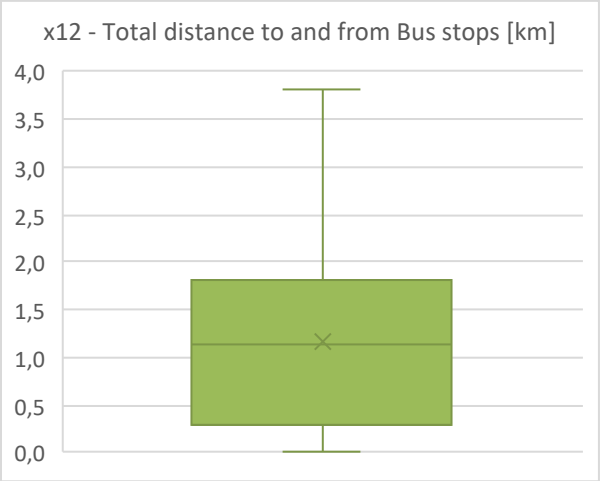

x905 - Annual average precipitation [zones 1-10]
12
10
8
6
4
2
0

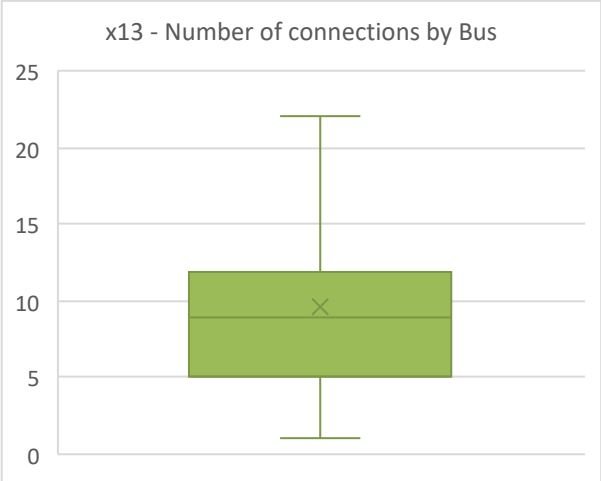

x915 - Annual average temperature [zones 1-8]
9
8
7
6
5
4
3
2
1
0

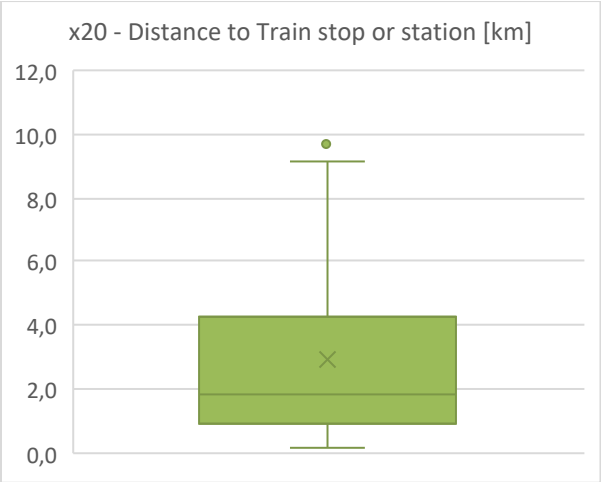
C.2 Variables related to connections of municipalities

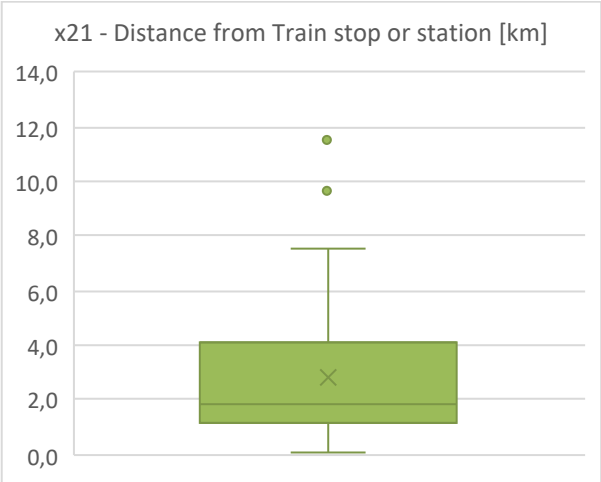


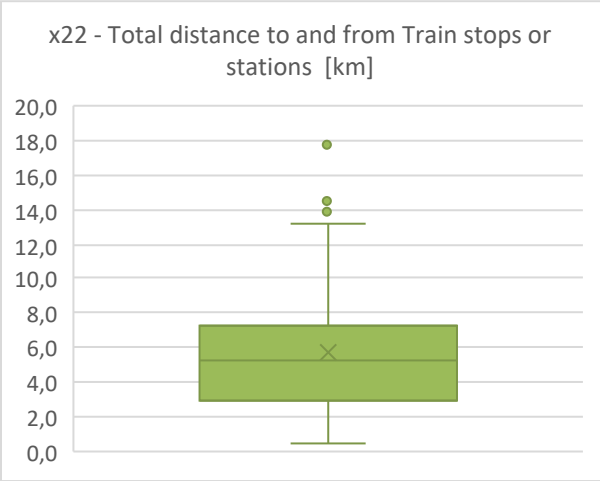


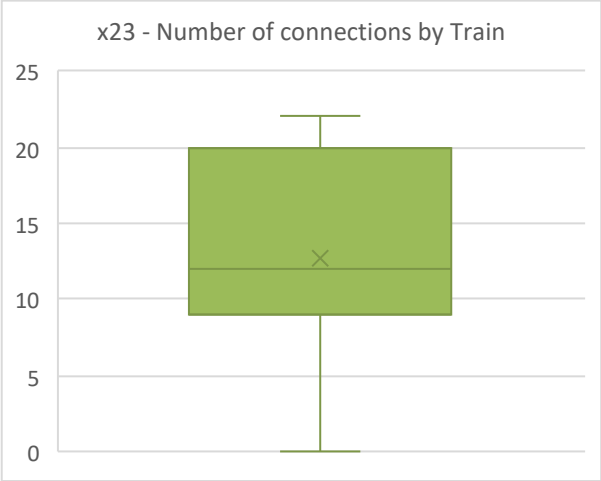


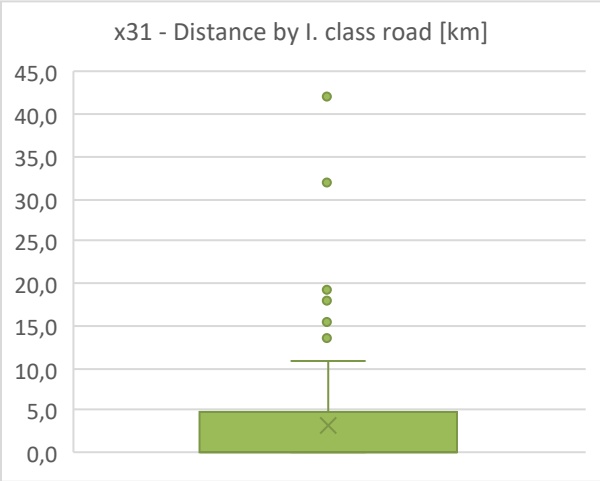


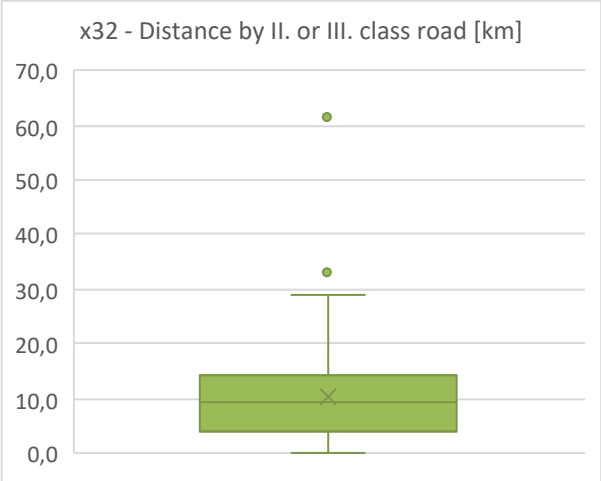


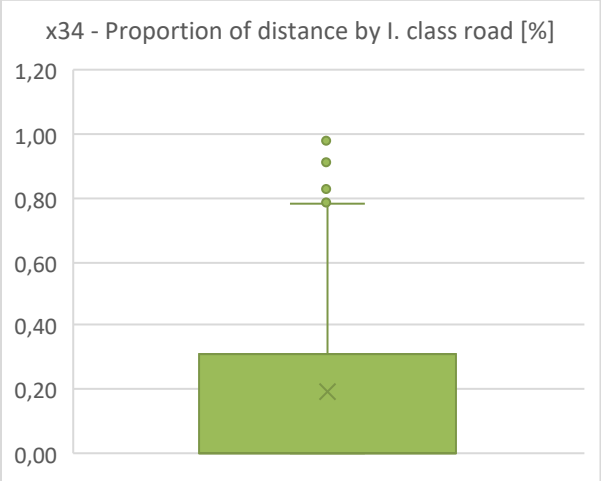


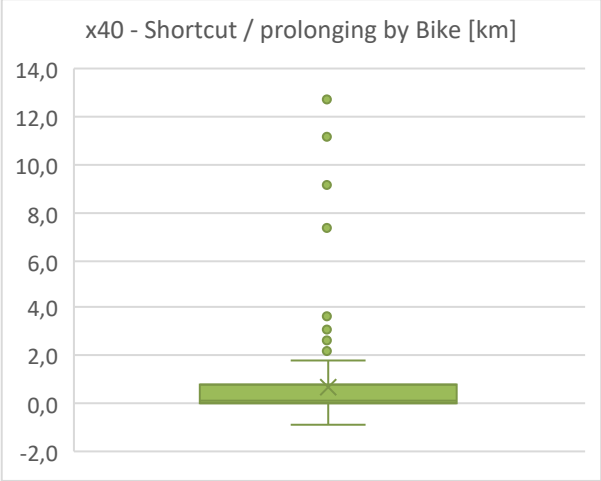

x22 - Total distance to and from Train stops or stations [km]
20,0
18,0
16,0
14,0
12,0
10,0
8,0
6,0
4,0
2,0
0,0

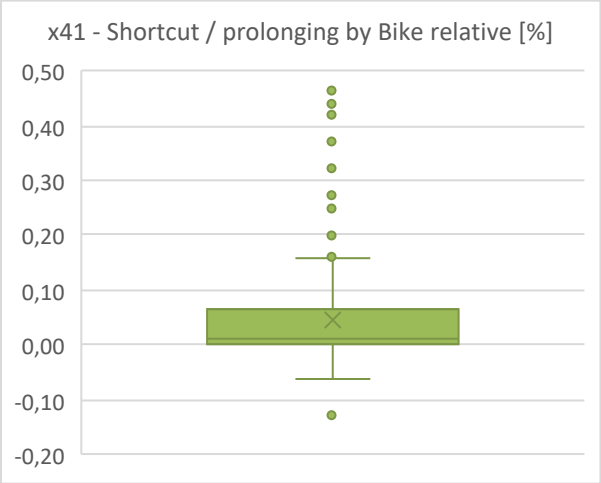

x23 - Number of connections by Train
25
20
15
10
5
0

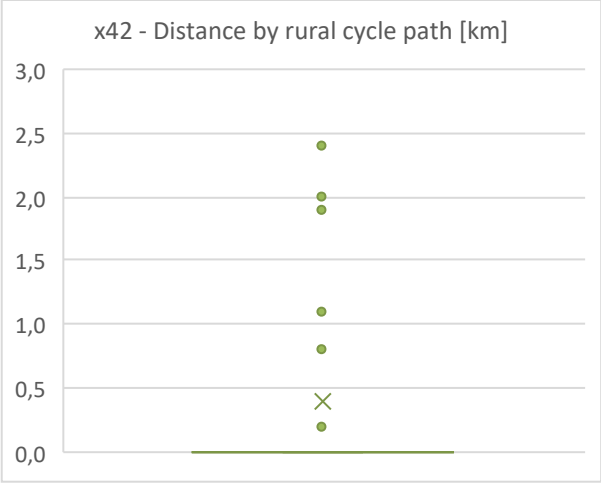

x31 - Distance by I. class road [km]
45,0
40,0
35,0
30,0
25,0
20,0
15,0
10,0
5,0
0,0

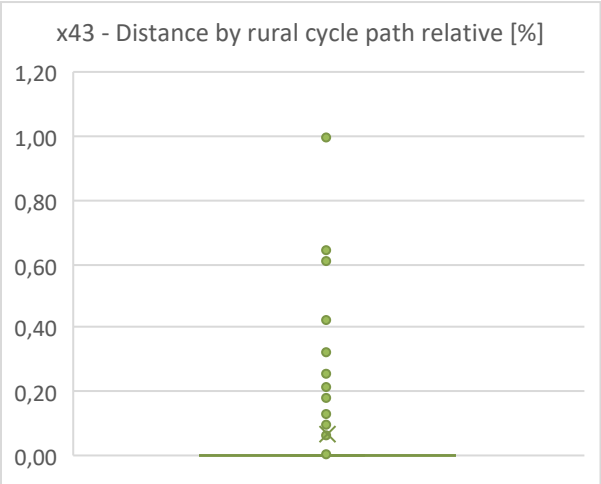

x32 - Distance by II. or III. class road [km]
70,0
60,0
50,0
40,0
30,0
20,0
10,0
0,0

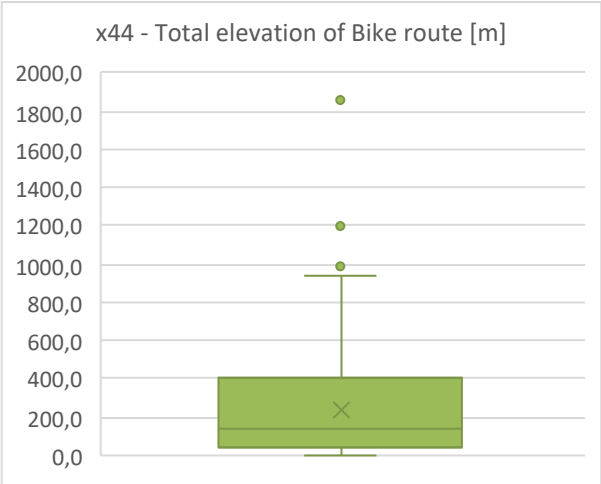

x34 - Proportion of distance by I. class road [%]
1,20
1,00
0,80
0,60
0,40
0,20
0,00

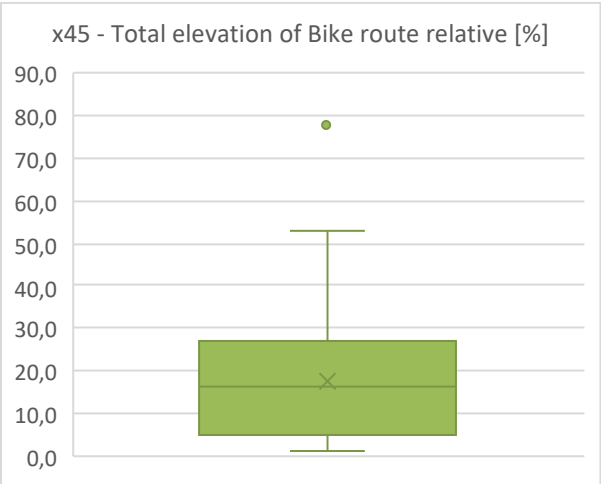

x40 - Shortcut / prolonging by Bike [km]
14,0
12,0
10,0
8,0
6,0
4,0
2,0
0,0
-2,0

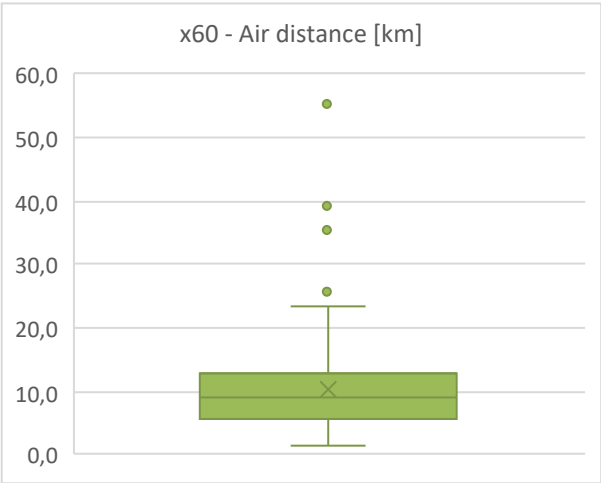

x41 - Shortcut / prolonging by Bike relative [%]
0,50
0,40
0,30
0,20
0,10
0,00
-0,10
-0,20

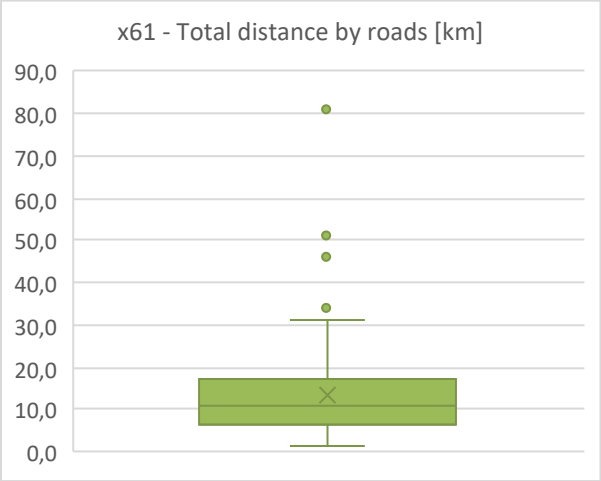

x42 - Distance by rural cycle path [km]
3,0
2,5
2,0
1,5
1,0
0,5
0,0

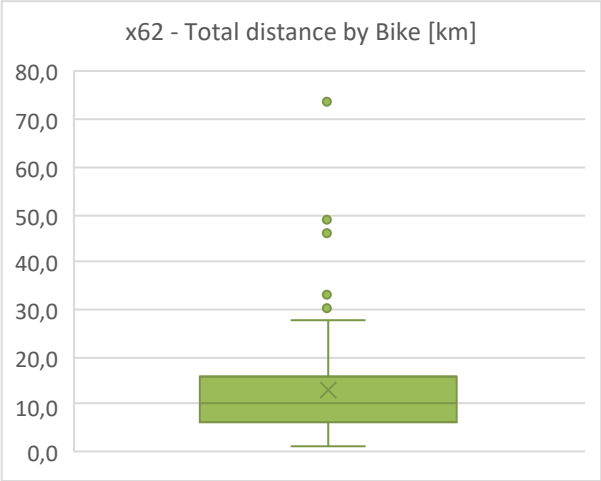

x43 - Distance by rural cycle path relative [%]
1,20
1,00
0,80
0,60
0,40
0,20
0,00

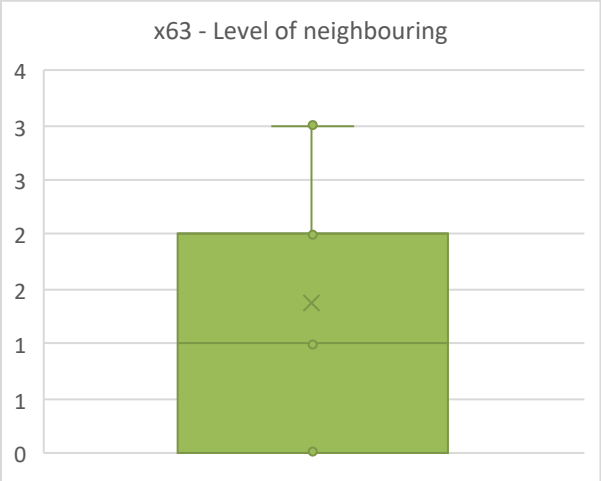

x44 - Total elevation of Bike route [m]
2000,0
1800,0
1600,0
1400,0
1200,0
1000,0
800,0
600,0
400,0
200,0
0,0

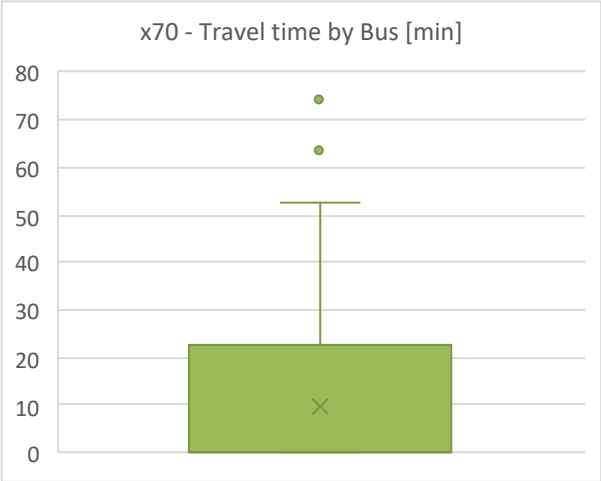

x45 - Total elevation of Bike route relative [%]
90,0
80,0
70,0
60,0
50,0
40,0
30,0
20,0
10,0
0,0

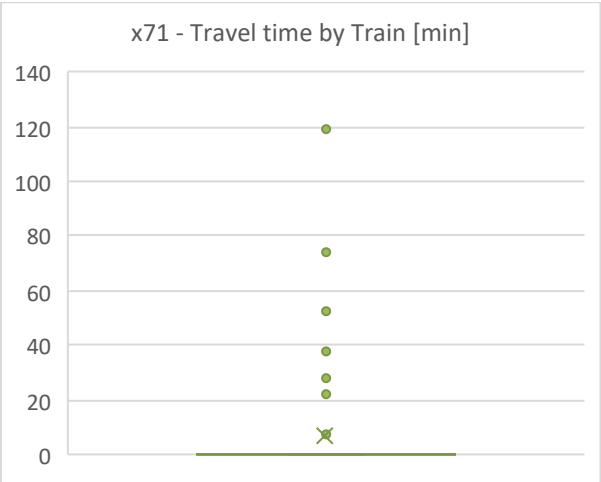

x60 - Air distance [km]
60,0
50,0
40,0
30,0
20,0
10,0
0,0

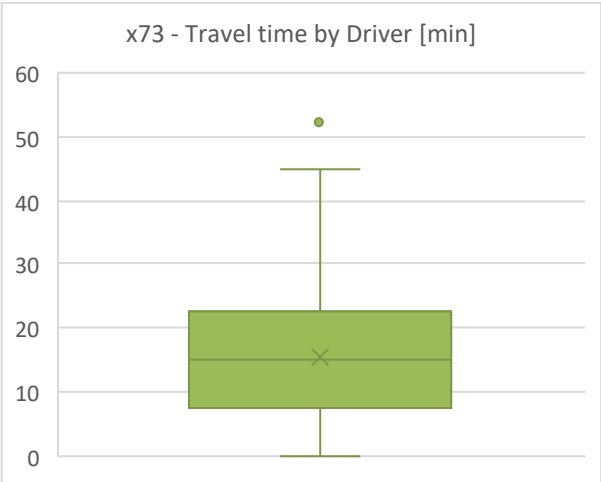

x61 - Total distance by roads [km]
90,0
80,0
70,0
60,0
50,0
40,0
30,0
20,0
10,0
0,0

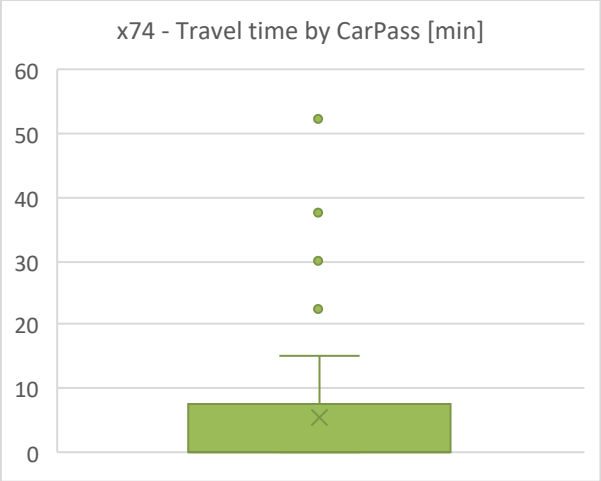

x62 - Total distance by Bike [km]
80,0
70,0
60,0
50,0
40,0
30,0
20,0
10,0
0,0

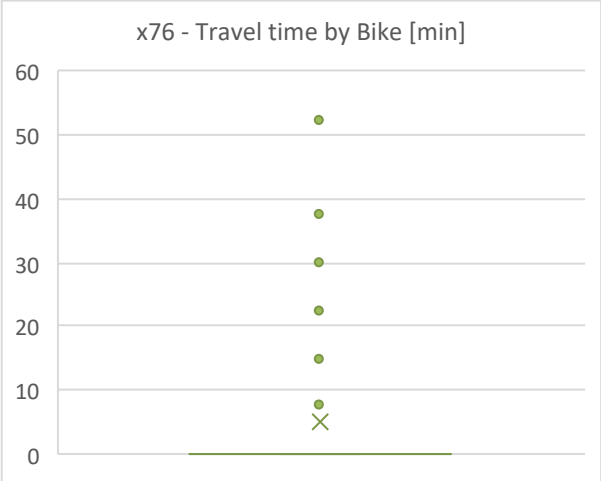

x63 - Level of neighbouring
4
3
3
2
2
1
1
0

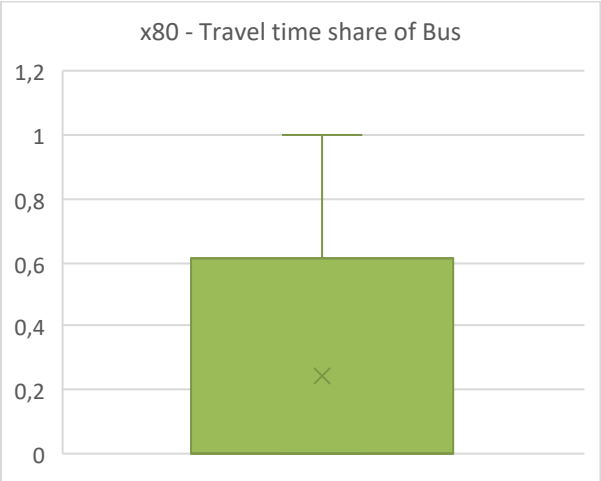

x70 - Travel time by Bus [min]
80
70
60
50
40
30
20
10
0


x71 - Travel time by Train [min]
140
120
100
80
60
40
20
0

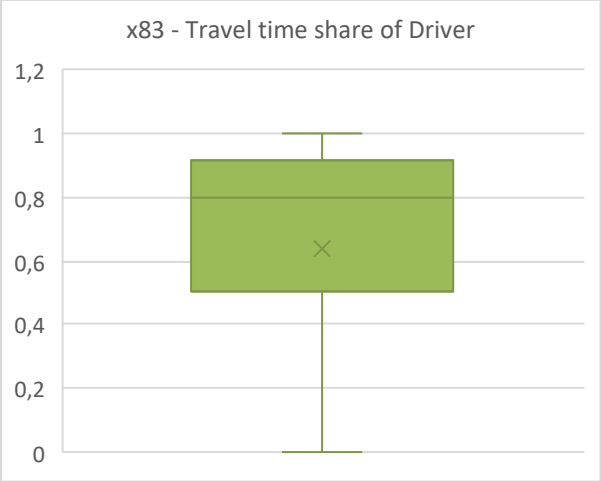

x73 - Travel time by Driver [min]
60
50
40
30
20
10
0

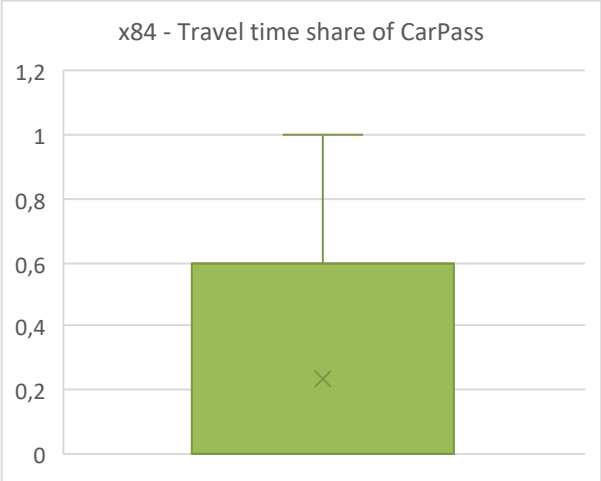

x74 - Travel time by CarPass [min]
60
50
40
30
20
10
0

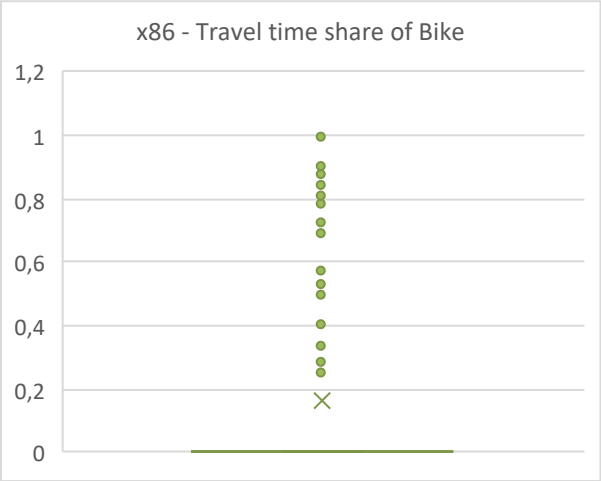

x76 - Travel time by Bike [min]
60
50
40
30
20
10
0

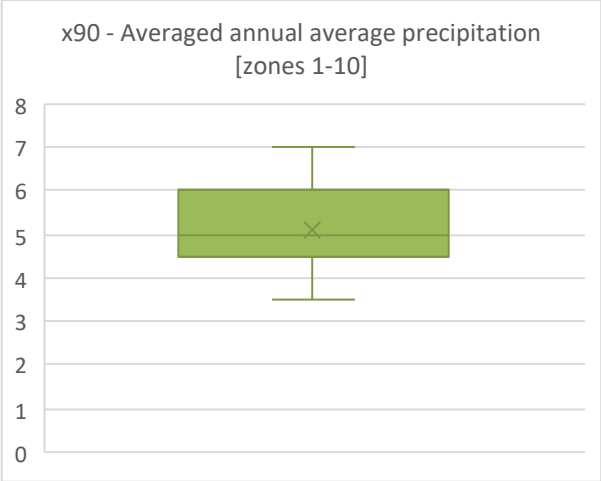

x80 - Travel time share of Bus
1,2
1
0,8
0,6
0,4
0,2
0

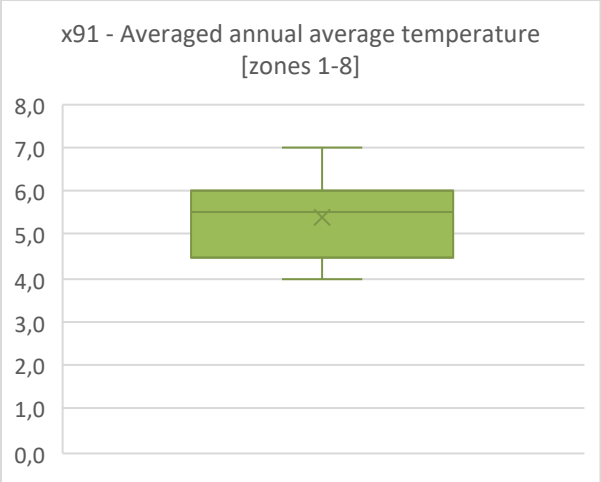

x81 - Travel time share of Train
1,2
1
0,8
0,6
0,4
0,2
0


x83 - Travel time share of Driver
1,2
1
0,8
0,6
0,4
0,2
0


x84 - Travel time share of CarPass
1,2
1
0,8
0,6
0,4
0,2
0


x86 - Travel time share of Bike
1,2
1
0,8
0,6
0,4
0,2
0

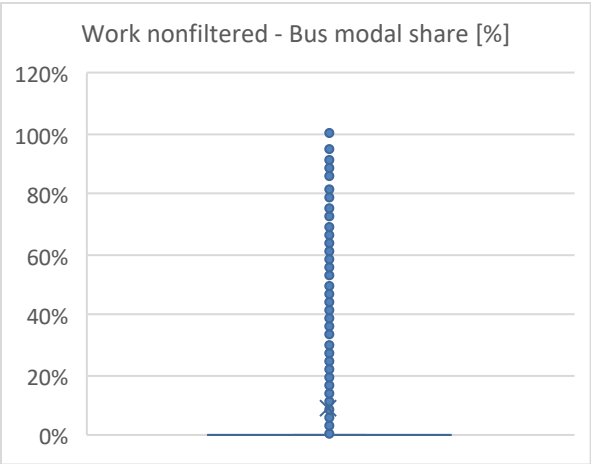

x90 - Averaged annual average precipitation [zones 1-10]
8
7
6
5
4
3
2
1
0

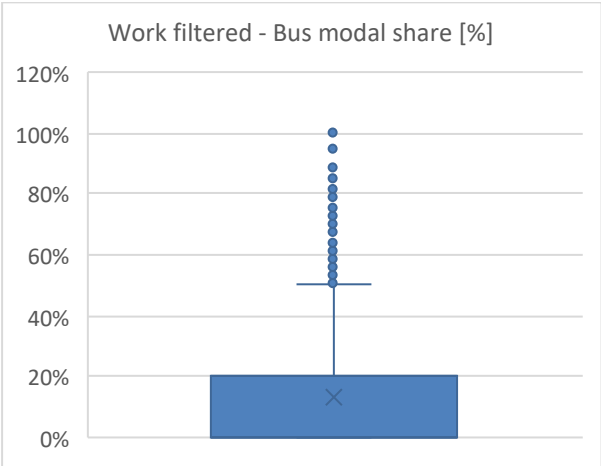

x91 - Averaged annual average temperature [zones 1-8]
8,0
7,0
6,0
5,0
4,0
3,0
2,0
1,0
0,0

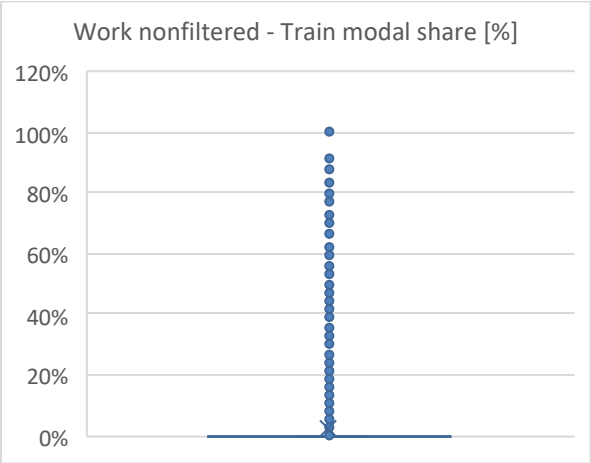
[D] Box and Whisker plots of explained variables

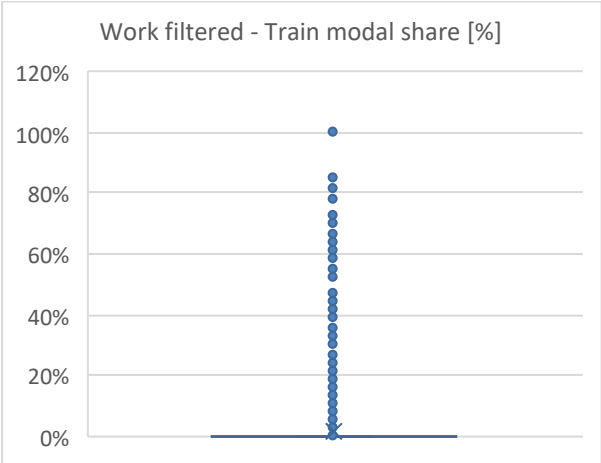
D.1 Modal shares of work commute

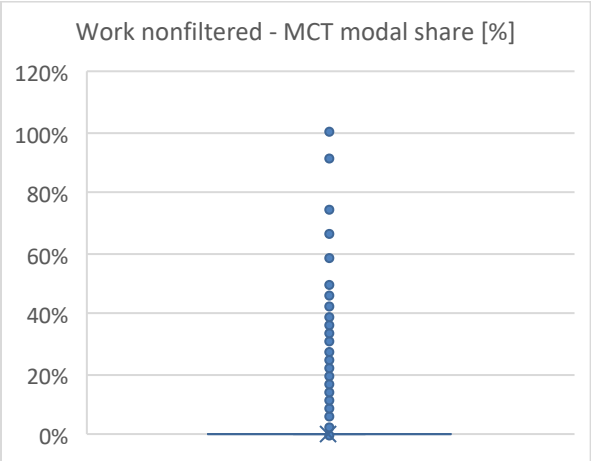
Nonfiltered data set of O-D pairs **Filtered data set of O-D pairs**

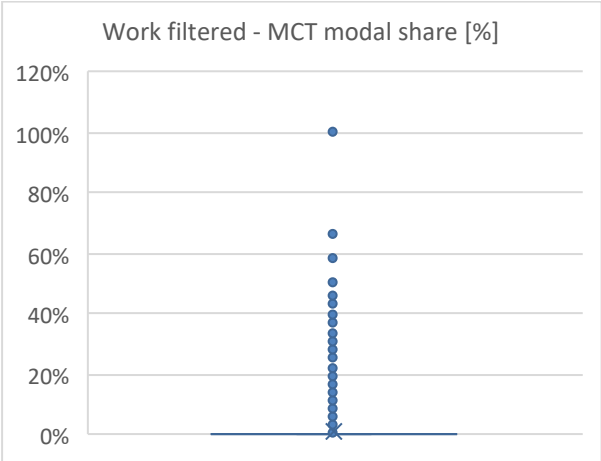




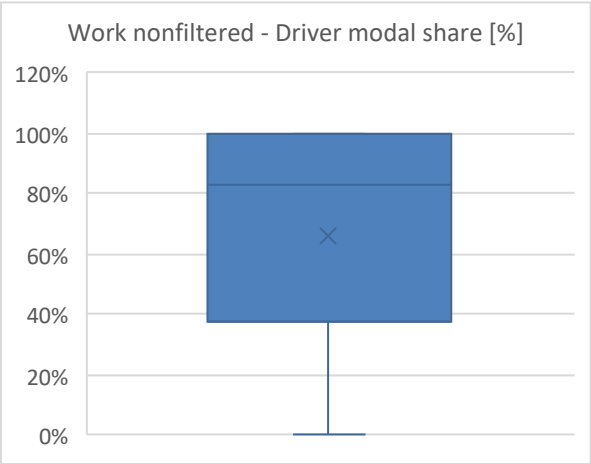


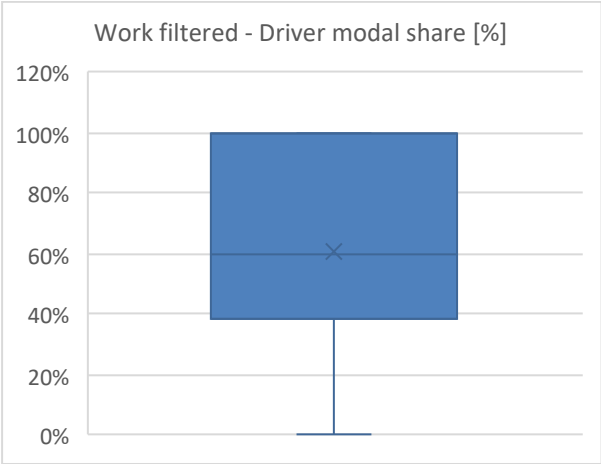


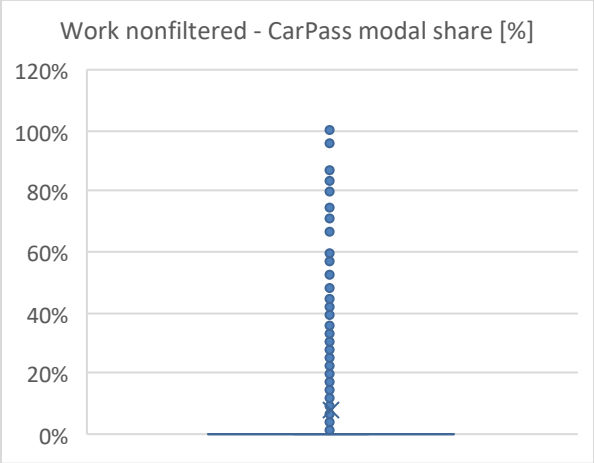


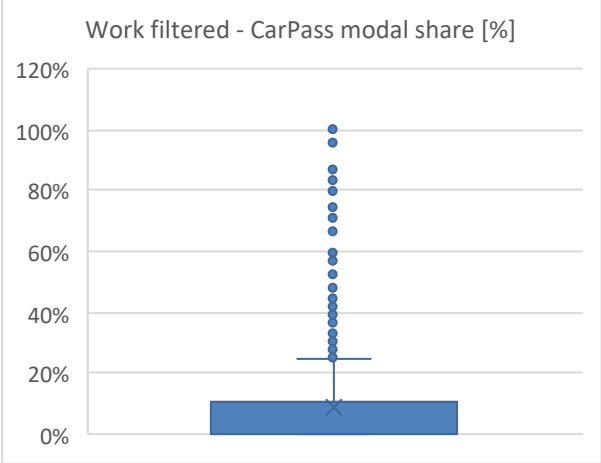


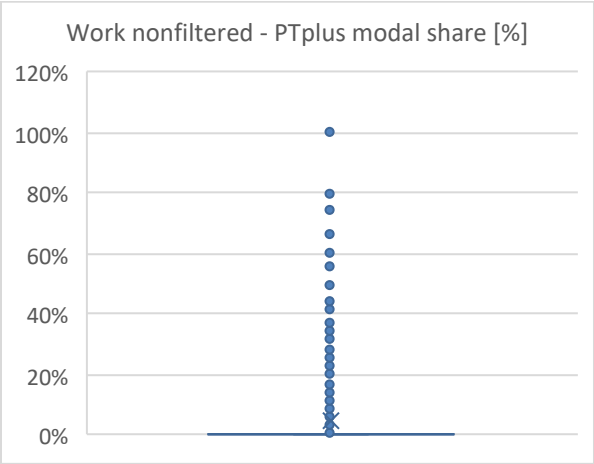
Nonfiltered data set of O-D pairs **Filtered data set of O-D pairs**

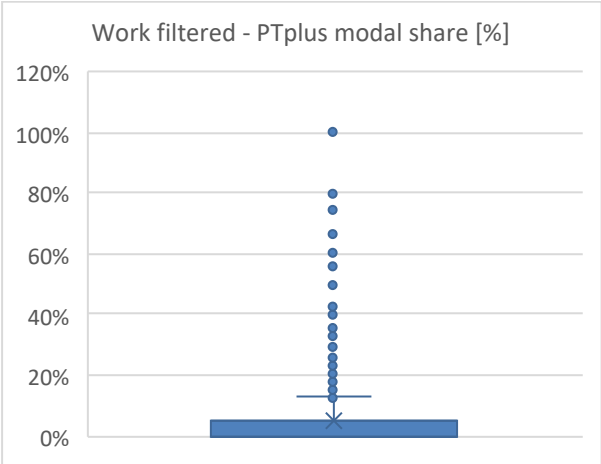




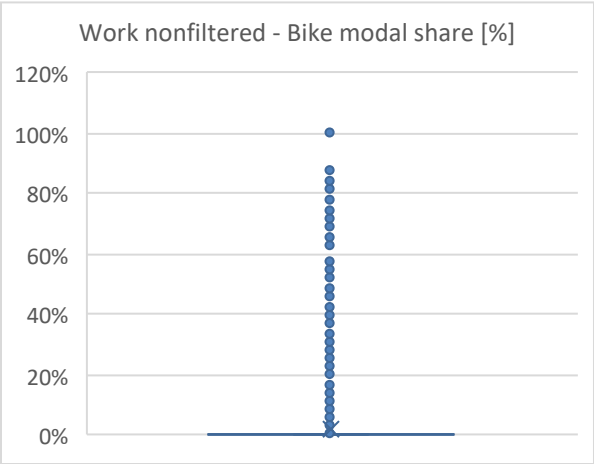


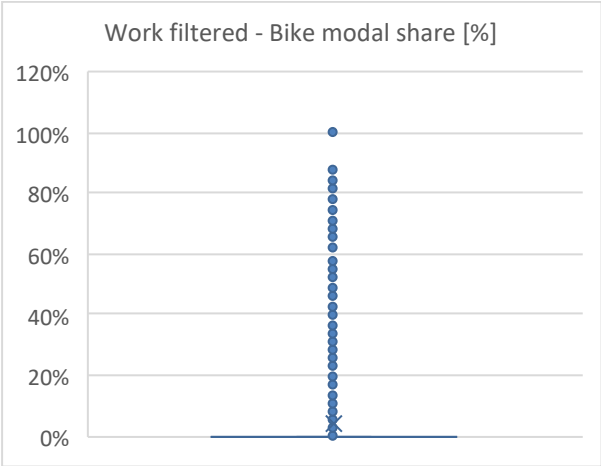


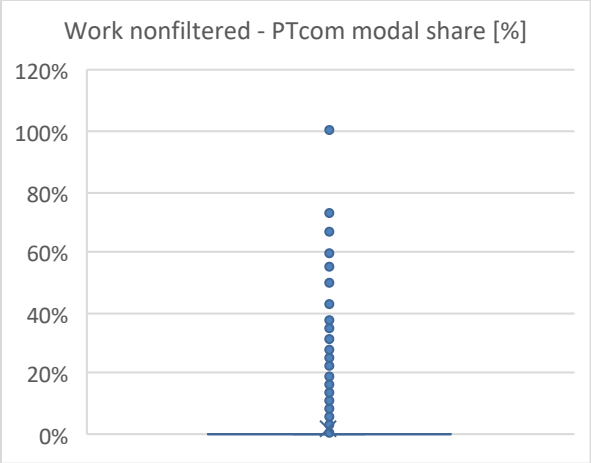


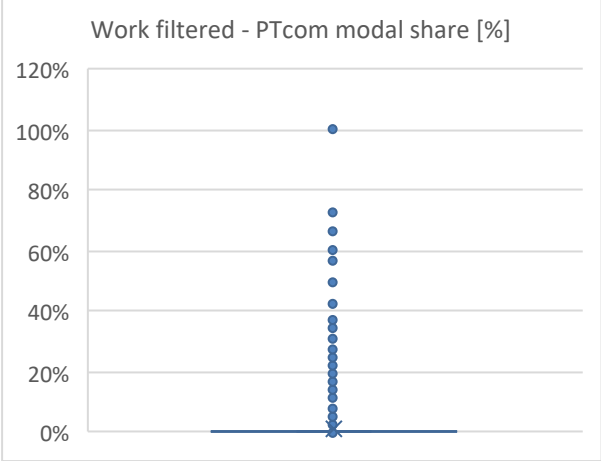


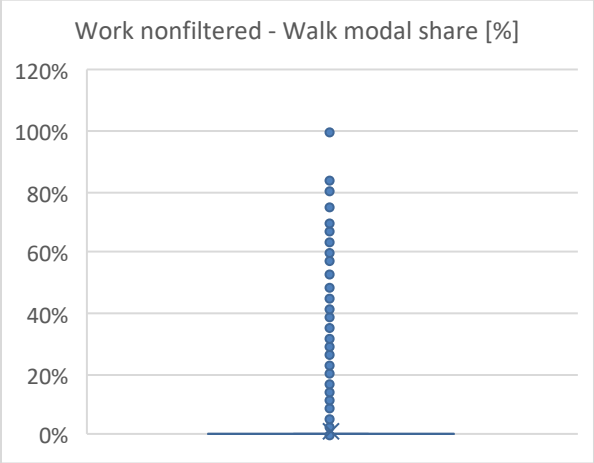
Nonfiltered data set of O-D pairs **Filtered data set of O-D pairs**

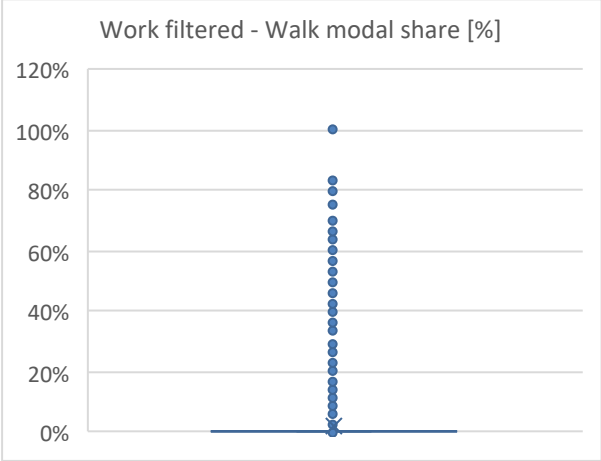






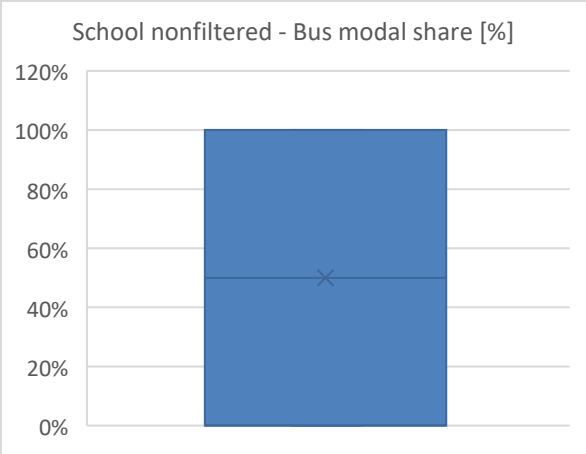


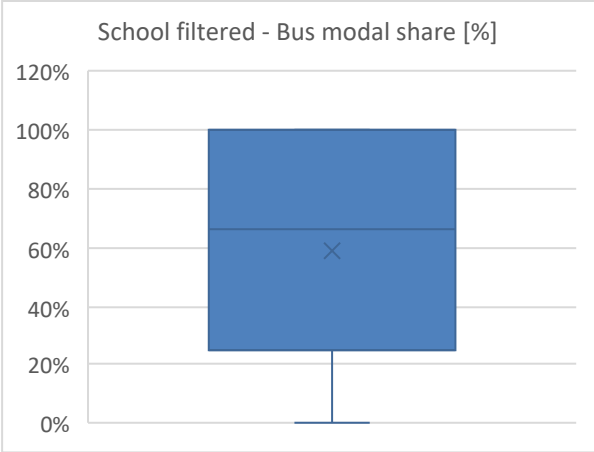


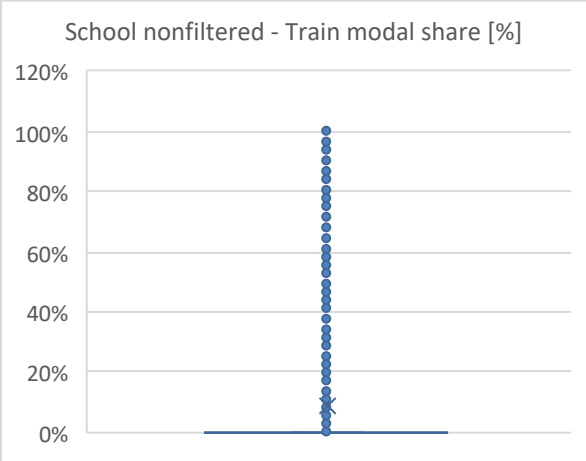


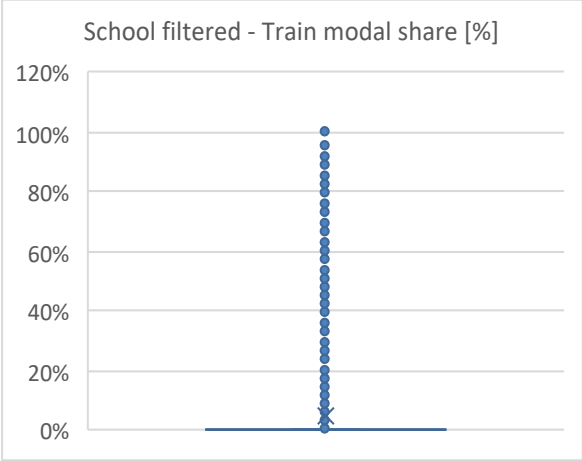
D.2 Modal shares of school commute

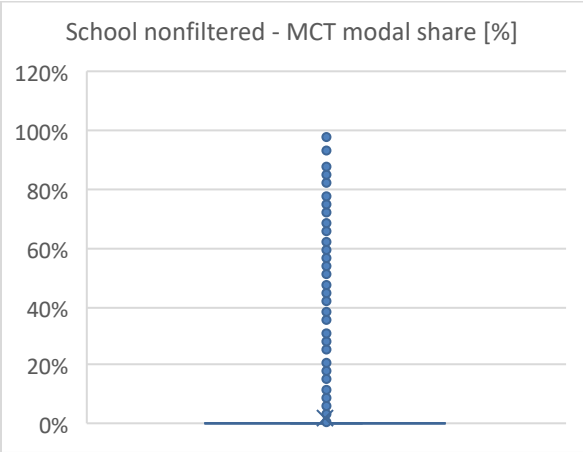
Nonfiltered data set of O-D pairs **Filtered data set of O-D pairs**

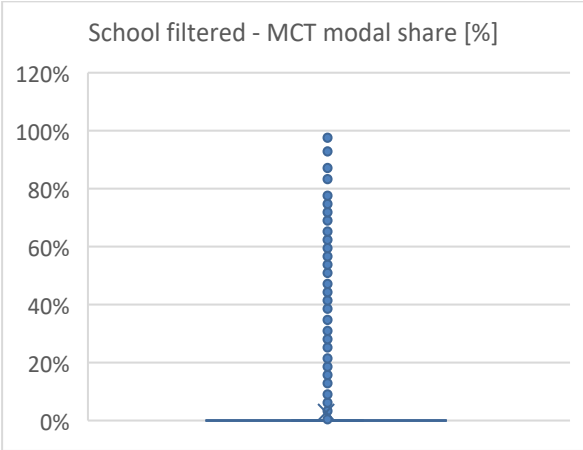




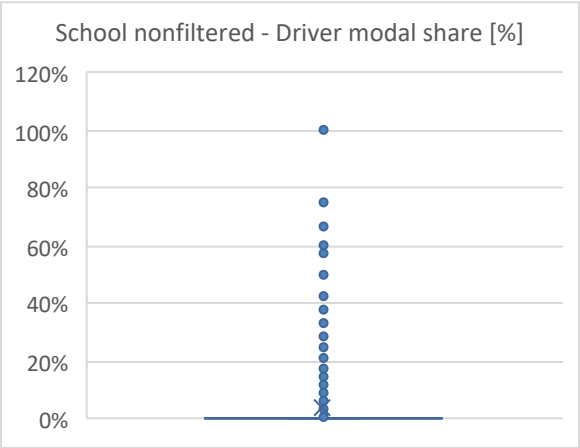


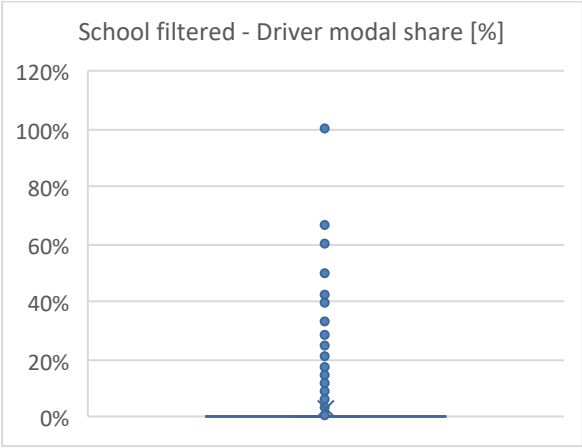


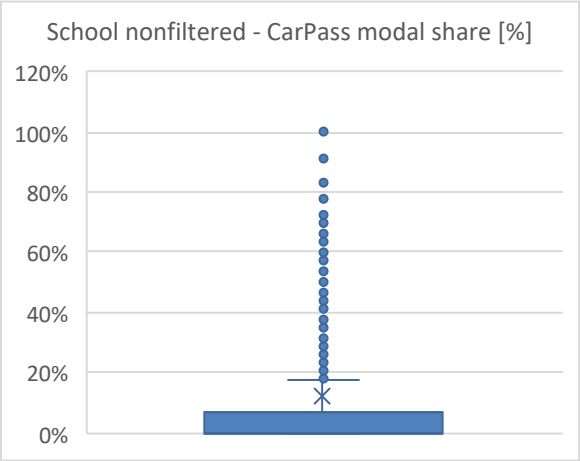


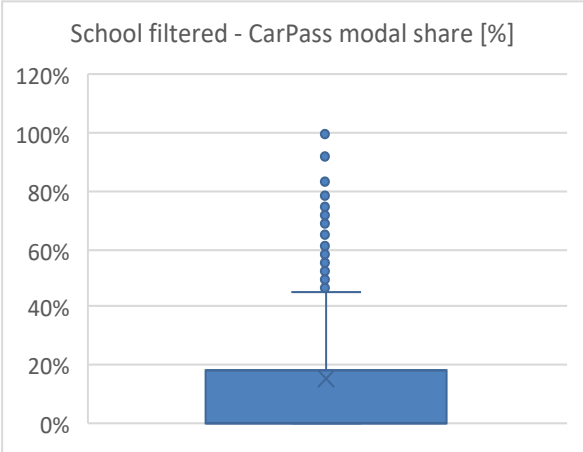


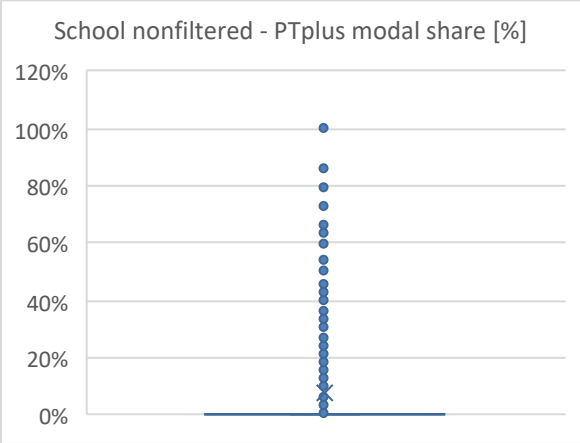
Nonfiltered data set of O-D pairs **Filtered data set of O-D pairs**

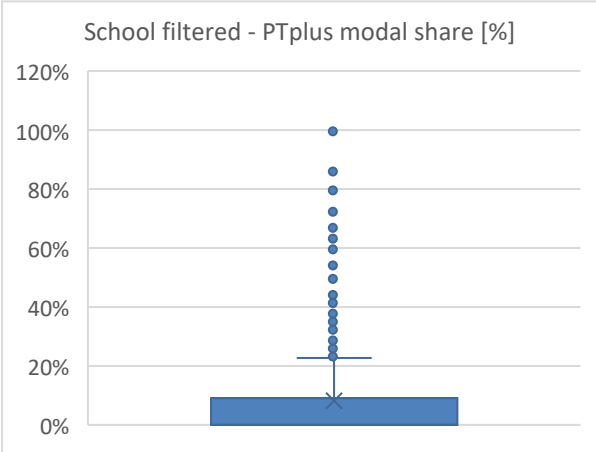




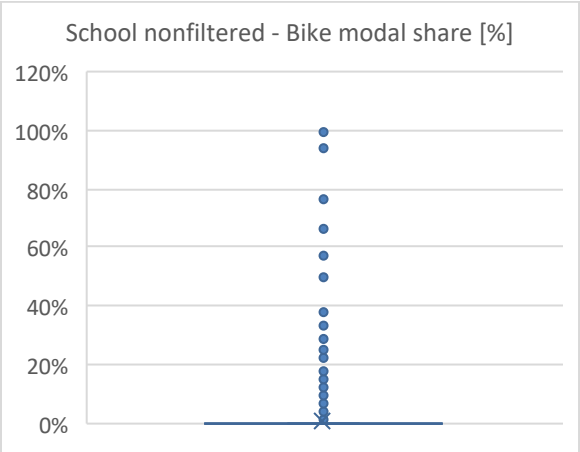


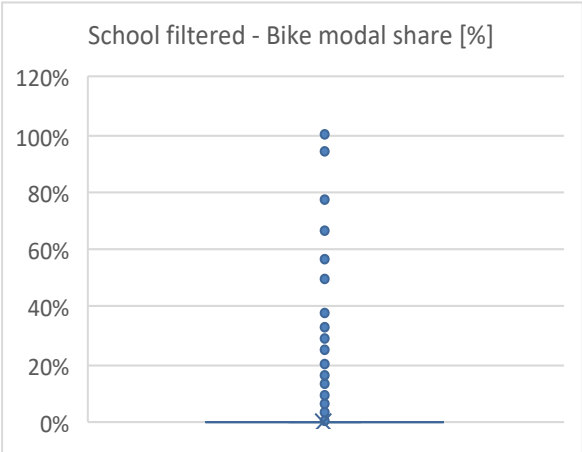


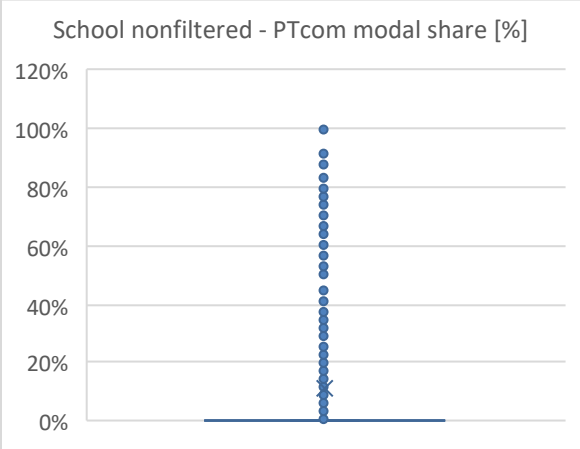


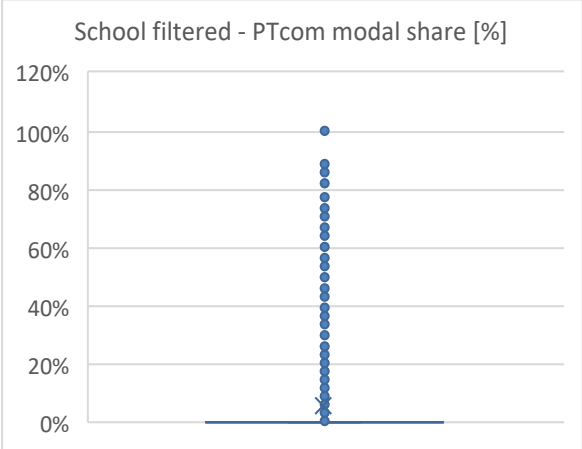


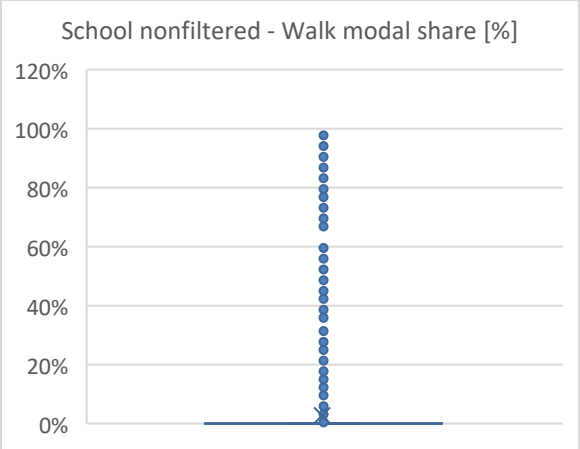
Nonfiltered data set of O-D pairs **Filtered data set of O-D pairs**

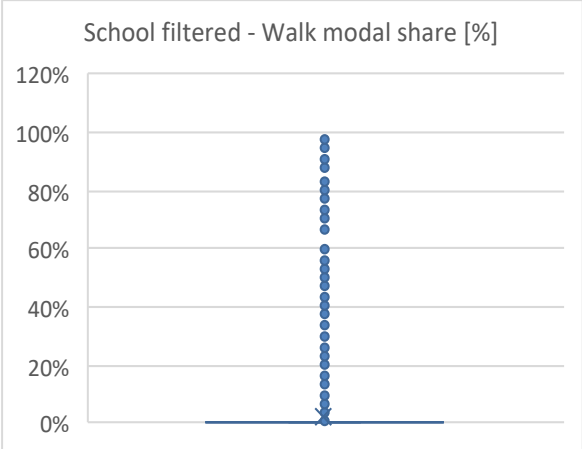












[E] Outcomes of linear regression models

Unless otherwise specified, outputs of model using **filtered** sets of **local** origin-destination pairs are presented in this chapter.

E.1 Type A models of school commute

E.1.1 Table of correlation coefficients – NONFILTERED data set

A-S-NF	x100	x110	x120	x130	x140	x200	x210	x220	x230	x240	x250	x260	x270	x280	x300	x310	x320	x330	x340	x350	x360	x370	x400	x410	x420	x430	x500	x510	x600	x610	x700	x710	x720	x730	x800	x810	x820	x905	x915
Bus	0.06	-0.05	-0.04	0.07	0.07	0.01	0.16	0.14	-0.11	-0.04	-0.02	-0.12	-0.14	-0.12	0.03	-0.05	0.09	0.02	0.05	0.00	-0.05	-0.14	0.07	0.06	-0.11	-0.11	-0.14	-0.16	-0.08	0.25	-0.12	-0.07	-0.12	0.14	0.09	-0.02	0.05	0.02	-0.08
Train	-0.01	0.01	0.01	-0.01	-0.02	-0.01	-0.05	-0.04	0.05	0.02	0.01	0.02	0.04	0.02	0.00	-0.01	0.00	0.01	0.00	-0.05	0.03	0.02	-0.06	-0.12	0.03	0.03	0.06	0.08	0.01	-0.25	0.01	0.03	0.02	-0.10	-0.07	-0.01	-0.05	0.00	0.05
MCT	-0.07	0.05	0.04	-0.07	-0.06	0.01	-0.10	-0.08	0.05	0.01	0.01	0.09	0.08	0.07	0.01	0.03	-0.08	-0.01	-0.03	0.01	0.02	0.12	-0.04	-0.02	0.07	0.07	0.11	0.12	0.06	-0.04	0.13	0.04	0.11	-0.09	-0.04	0.02	0.01	-0.02	0.03
Driver	0.00	-0.01	0.03	-0.01	0.00	0.00	-0.02	-0.02	0.01	0.01	0.00	0.02	0.02	0.01	-0.02	0.02	-0.01	-0.01	-0.02	0.00	0.03	0.00	0.00	0.02	0.04	0.03	0.02	0.02	0.02	0.00	0.01	0.00	0.02	-0.01	0.01	0.02	-0.03	0.00	0.00
CarPass	-0.01	0.03	-0.01	-0.02	-0.01	-0.01	-0.09	-0.08	0.05	0.02	0.02	0.08	0.08	0.07	-0.04	0.03	-0.04	-0.06	-0.03	0.07	0.00	0.09	0.03	0.08	0.02	0.02	0.02	0.00	0.03	-0.01	0.10	0.06	0.08	0.04	0.01	0.01	0.01	-0.01	0.00
PTplus	0.00	0.01	-0.01	0.00	0.00	-0.01	0.00	-0.01	0.01	0.00	0.01	-0.01	0.00	0.00	-0.03	0.00	-0.01	0.00	0.01	0.03	-0.02	-0.01	0.00	0.05	-0.03	-0.03	-0.03	-0.04	-0.02	0.00	-0.01	-0.02	-0.03	0.04	-0.01	-0.03	0.00	0.01	0.00
Bike	0.00	0.00	0.01	-0.01	-0.01	-0.02	0.00	-0.01	0.01	0.00	0.01	0.00	0.01	0.01	-0.02	0.01	-0.01	0.00	-0.01	0.00	0.00	0.00	-0.01	0.01	0.00	0.00	0.00	0.00	-0.01	-0.02	0.01	0.01	0.00	-0.01	-0.02	-0.02	-0.01	-0.01	0.02
PTcom	-0.04	0.01	0.03	-0.04	-0.03	0.01	-0.04	-0.03	0.02	0.00	-0.02	0.03	0.03	0.02	0.03	0.02	-0.04	0.02	-0.03	-0.04	0.01	0.04	-0.06	-0.09	0.03	0.03	0.06	0.09	0.01	-0.09	0.01	0.01	0.02	-0.11	-0.08	0.01	-0.02	-0.03	0.07
Walk	-0.01	0.00	0.02	-0.02	-0.01	0.00	-0.05	-0.04	0.02	0.02	0.01	0.04	0.04	0.03	0.01	0.01	-0.02	0.01	-0.01	-0.02	0.06	0.03	0.00	-0.05	0.10	0.10	0.09	0.10	0.07	-0.05	0.02	0.02	0.06	-0.07	0.02	0.04	-0.02	0.02	-0.02

E.1.2 Table of correlation coefficients

A-S-F	x100	x110	x120	x130	x140	x200	x210	x220	x230	x240	x250	x260	x270	x280	x300	x310	x320	x330	x340	x350	x360	x370	x400	x410	x420	x430	x500	x510	x600	x610	x700	x710	x720	x730	x800	x810	x820	x905	x915
Bus	0.06	-0.05	-0.04	0.08	0.06	0.02	0.22	0.20	-0.13	-0.05	-0.05	-0.19	-0.20	-0.18	0.05	-0.07	0.11	0.05	0.06	-0.03	-0.07	-0.20	0.04	0.01	-0.14	-0.14	-0.18	-0.18	-0.11	0.26	-0.18	-0.10	-0.18	0.12	0.07	-0.02	0.03	0.02	-0.08
Train	-0.01	0.00	0.02	-0.02	-0.02	0.00	-0.06	-0.06	0.05	0.01	0.01	0.03	0.05	0.04	0.00	0.01	-0.01	-0.01	-0.01	-0.04	0.04	0.05	-0.05	-0.10	0.02	0.03	0.06	0.08	0.01	-0.25	0.03	0.04	0.04	-0.09	-0.05	0.00	-0.05	-0.01	0.05
MCT	-0.11	0.08	0.06	-0.11	-0.08	0.00	-0.16	-0.13	0.08	0.02	0.02	0.14	0.13	0.12	0.00	0.06	-0.13	0.00	-0.05	0.03	0.04	0.19	-0.05	-0.02	0.12	0.12	0.18	0.19	0.10	-0.07	0.20	0.09	0.19	-0.13	-0.07	0.02	0.01	-0.05	0.06
Driver	0.01	-0.02	0.02	0.00	0.01	0.00	-0.04	-0.04	0.02	0.02	0.01	0.04	0.04	0.04	-0.02	0.02	-0.01	-0.02	-0.02	-0.01	0.03	0.02	0.01	0.03	0.08	0.08	0.06	0.04	0.06	-0.02	0.03	0.02	0.05	-0.02	0.00	0.02	-0.04	0.00	0.01
CarPass	-0.02	0.04	0.00	-0.03	-0.03	-0.01	-0.12	-0.10	0.07	0.02	0.02	0.09	0.10	0.09	-0.05	0.04	-0.05	-0.06	-0.04	0.07	0.01	0.10	0.02	0.07	0.03	0.04	0.03	0.02	0.03	-0.06	0.10	0.05	0.09	0.02	0.01	0.02	0.01	0.00	0.00
PTplus	0.00	0.01	-0.02	0.01	0.00	-0.02	-0.02	-0.02	0.02	0.00	0.02	0.02	0.02	0.02	-0.04	0.00	-0.01	-0.01	0.01	0.03	-0.02	0.00	0.00	0.04	-0.02	-0.02	-0.02	-0.03	-0.01	-0.01	-0.01	-0.02	-0.02	0.04	-0.01	-0.02	0.00	0.01	0.00
Bike	-0.01	0.00	0.01	-0.01	-0.01	-0.02	0.00	-0.01	0.01	0.00	0.01	0.00	0.01	0.01	-0.02	0.02	-0.01	0.01	-0.02	0.00	0.01	0.00	-0.02	0.00	0.00	0.00	0.01	0.01	-0.01	-0.04	0.00	0.02	0.00	-0.02	-0.04	-0.03	-0.03	-0.01	0.03
PTcom	-0.02	0.00	0.03	-0.03	-0.01	0.01	-0.04	-0.03	0.01	0.02	-0.01	0.04	0.04	0.03	0.03	0.01	-0.03	-0.01	-0.01	-0.02	0.03	0.05	-0.03	-0.06	0.05	0.05	0.07	0.08	0.03	-0.11	0.02	0.02	0.03	-0.08	-0.04	0.02	-0.02	-0.02	0.05
Walk	-0.02	0.01	0.01	-0.02	-0.01	-0.03	-0.07	-0.07	0.04	0.02	0.02	0.07	0.07	0.07	-0.01	0.02	-0.03	0.01	-0.02	-0.01	0.04	0.04	-0.01	-0.04	0.10	0.10	0.10	0.10	0.07	-0.05	0.05	0.05	0.07	-0.06	-0.02	0.02	0.01	0.00	0.02

E.1.3 Table of correlation coefficients – data set with ONLY ORIGIN-DESTINATION PAIRS USED IN TYPE C MODEL of school commute

A-S-F ŠTBK	x100	x110	x120	x130	x140	x200	x210	x220	x230	x240	x250	x260	x270	x280	x300	x310	x320	x330	x340	x350	x360	x370	x400	x410	x420	x430	x500	x510	x600	x610	x700	x710	x720	x730	x800	x810	x820	x905	x915
Bus	-0.11	0.05	-0.02	-0.03	0.04	0.32	0.32	0.40	-0.31	-0.04	-0.18	-0.32	-0.38	-0.37	0.16	-0.08	-0.13	0.26	0.07	-0.10	-0.16	-0.15	-0.04	-0.14	-0.17	-0.18	0.00	-0.06	0.00	0.18	-0.02	-0.03	-0.07	-0.06	0.16	-0.09	-0.06	0.11	-0.18
Train	-0.11	0.09	0.01	-0.11	-0.13	-0.15	-0.15	-0.19	0.22	-0.09	0.10	0.09	0.17	0.13	-0.21	0.10	-0.10	-0.03	-0.10	-0.11	0.03	0.08	-0.20	-0.12	-0.07	-0.06	-0.02	0.00	0.00	-0.36	0.07	0.07	0.02	-0.11	-0.28	-0.25	-0.14	-0.30	0.15
Driver	0.18	-0.01	-0.36	0.27	0.07	-0.05	-0.11	-0.12	-0.28	-0.33	-0.17	0.47	0.13	0.35	0.20	-0.34	0.22	0.21	0.38	0.73	-0.07	-0.08	0.35	0.12	-0.06	-0.07	-0.11	-0.10	0.00	0.16	0.02	0.03	-0.01	0.09	0.14	0.11	0.30	0.14	-0.11
CarPass	0.16	-0.11	0.11	0.03	0.05	-0.07	-0.01	-0.04	-0.02	0.14	0.19	0.02	0.04	0.05	-0.02	-0.05	0.22	-0.39	0.04	-0.04	-0.03	0.02	0.14	0.25	0.07	0.07	-0.08	-0.07	0.00	0.07	-0.07	-0.07	-0.01	0.25	0.10	0.23	0.12	0.14	0.01

E.1.4 Table of coefficients of determination – NONFILTERED data set

A-S-NF	x100	x110	x120	x130	x140	x200	x210	x220	x230	x240	x250	x260	x270	x280	x300	x310	x320	x330	x340	x350	x360	x370	x400	x410	x420	x430	x500	x510	x600	x610	x700	x710	x720	x730	x800	x810	x820	x905	x915	AVERAGE
Bus	0.003	0.002	0.002	0.005	0.004	5E-05	0.027	0.020	0.012	0.001	5E-04	0.015	0.020	0.013	9E-04	0.003	0.009	6E-04	0.003	2E-05	0.003	0.019	0.004	0.004	0.012	0.012	0.021	0.027	0.006	0.061	0.014	0.005	0.015	0.021	0.008	4E-04	0.003	4E-04	0.007	**0.00983**
Train	1E-04	7E-05	4E-05	1E-04	6E-04	7E-05	0.002	0.002	0.003	3E-04	6E-05	3E-04	0.002	5E-04	1E-06	4E-05	2E-05	7E-05	8E-06	0.003	0.001	4E-04	0.003	0.014	7E-04	8E-04	0.003	0.007	1E-04	0.065	9E-05	0.001	4E-04	0.011	0.005	4E-05	0.003	1E-06	0.002	**0.00339**
MCT	0.005	0.003	0.002	0.006	0.003	1E-04	0.010	0.007	0.002	2E-04	6E-05	0.008	0.006	0.005	1E-04	0.001	0.007	7E-05	8E-04	2E-04	5E-04	0.015	0.001	6E-04	0.005	0.005	0.012	0.014	0.003	0.002	0.016	0.002	0.013	0.009	0.002	6E-04	7E-05	6E-04	0.001	**0.00431**
Driver	5E-06	2E-04	6E-04	2E-04	1E-05	2E-05	4E-04	2E-04	1E-04	3E-05	1E-05	2E-04	3E-04	2E-04	3E-04	3E-04	2E-04	1E-04	3E-04	1E-05	7E-04	2E-05	1E-06	6E-04	0.001	0.001	5E-04	4E-04	4E-04	9E-06	5E-05	1E-05	3E-04	8E-05	4E-05	2E-04	8E-04	7E-06	1E-05	**0.00026**
CarPass	4E-05	0.001	1E-04	3E-04	1E-04	9E-05	0.008	0.006	0.002	2E-04	6E-04	0.006	0.006	0.005	0.002	0.001	0.001	0.003	7E-04	0.005	1E-08	0.008	7E-04	0.007	6E-04	6E-04	4E-04	2E-05	1E-03	2E-04	0.010	0.003	0.007	0.001	2E-04	3E-05	2E-04	1E-04	5E-06	**0.00227**
PTplus	7E-08	1E-04	1E-04	2E-05	5E-06	2E-04	7E-09	3E-05	2E-04	1E-06	8E-05	4E-05	2E-05	5E-07	0.001	5E-06	3E-05	4E-07	4E-05	7E-04	5E-04	1E-04	2E-08	0.003	0.001	0.001	0.001	0.002	5E-04	2E-05	1E-04	5E-04	8E-04	0.002	1E-04	7E-04	2E-06	5E-05	1E-05	**0.00041**
Bike	2E-05	2E-05	3E-05	5E-05	6E-05	3E-04	2E-05	1E-04	2E-04	1E-05	3E-05	4E-06	1E-04	6E-05	2E-04	2E-04	5E-05	2E-05	2E-04	3E-06	1E-06	6E-06	1E-04	8E-05	4E-08	3E-08	2E-05	9E-06	4E-05	4E-04	3E-05	2E-04	3E-08	1E-04	4E-04	4E-04	2E-04	8E-05	4E-04	**0.00011**
PTcom	0.001	1E-04	0.001	0.001	0.001	9E-05	0.001	7E-04	4E-04	5E-06	3E-04	9E-04	7E-04	5E-04	7E-04	6E-04	0.002	2E-04	7E-04	0.001	7E-05	0.001	0.003	0.007	8E-04	8E-04	0.003	0.008	1E-04	0.008	2E-04	2E-04	3E-04	0.012	0.006	1E-04	3E-04	9E-04	0.005	**0.00186**
Walk	2E-04	1E-06	3E-04	3E-04	2E-04	2E-05	0.002	0.001	4E-04	3E-04	8E-05	0.001	0.001	0.001	1E-04	4E-05	3E-04	5E-05	1E-04	5E-04	0.004	6E-04	4E-07	0.002	0.010	0.010	0.008	0.009	0.006	0.002	5E-04	2E-04	0.003	0.004	3E-04	0.002	6E-04	6E-04	2E-04	**0.00189**
AVG	**0.001**	**7E-04**	**6E-04**	**0.001**	**0.001**	**1E-04**	**0.006**	**0.004**	**0.002**	**3E-04**	**2E-04**	**0.003**	**0.004**	**0.003**	**6E-04**	**7E-04**	**0.002**	**5E-04**	**6E-04**	**0.001**	**0.001**	**0.005**	**0.001**	**0.004**	**0.003**	**0.003**	**0.005**	**0.008**	**0.002**	**0.015**	**0.005**	**0.001**	**0.004**	**0.007**	**0.002**	**5E-04**	**9E-04**	**3E-04**	**0.002**	0.00270

E.1.5 Table of coefficients of determination

A-S-F	x100	x110	x120	x130	x140	x200	x210	x220	x230	x240	x250	x260	x270	x280	x300	x310	x320	x330	x340	x350	x360	x370	x400	x410	x420	x430	x500	x510	x600	x610	x700	x710	x720	x730	x800	x810	x820	x905	x915	AVERAGE
Bus	0.004	0.003	0.002	0.006	0.004	4E-04	0.050	0.040	0.018	0.002	0.002	0.035	0.039	0.031	0.003	0.005	0.012	0.003	0.004	0.001	0.005	0.039	0.002	1E-04	0.019	0.020	0.032	0.032	0.012	0.066	0.031	0.010	0.031	0.014	0.006	6E-04	0.001	6E-04	0.006	**0.01510**
Train	8E-05	3E-07	3E-04	2E-04	5E-04	2E-05	0.004	0.003	0.003	2E-04	2E-04	0.001	0.003	0.001	2E-06	4E-05	2E-04	3E-05	2E-04	0.002	0.001	0.003	0.003	0.010	6E-04	7E-04	0.004	0.007	1E-04	0.063	0.001	0.002	0.002	0.008	0.003	7E-08	0.002	1E-04	0.002	**0.00338**
MCT	0.012	0.006	0.004	0.013	0.007	2E-05	0.024	0.018	0.006	5E-04	5E-04	0.020	0.017	0.015	1E-05	0.004	0.017	2E-06	0.003	7E-04	0.002	0.034	0.002	6E-04	0.014	0.014	0.034	0.036	0.010	0.004	0.038	0.008	0.035	0.017	0.006	3E-04	3E-05	0.002	0.004	**0.01096**
Driver	1E-04	5E-04	3E-04	6E-06	5E-05	2E-06	0.002	0.002	4E-04	3E-04	2E-04	0.002	0.002	0.001	4E-04	5E-04	1E-04	4E-04	4E-04	4E-05	0.001	5E-04	3E-05	1E-03	0.006	0.006	0.003	0.001	0.004	3E-04	9E-04	6E-04	0.003	5E-04	3E-08	2E-04	0.001	2E-05	1E-04	**0.00108**
CarPass	3E-04	0.002	4E-06	0.001	8E-04	3E-05	0.013	0.010	0.005	2E-04	5E-04	0.009	0.010	0.007	0.002	0.002	0.002	0.003	0.001	0.004	4E-05	0.010	3E-04	0.006	0.001	0.001	0.001	4E-04	0.001	0.003	0.010	0.002	0.007	2E-04	1E-04	4E-04	2E-04	3E-06	8E-06	**0.00304**
PTplus	3E-06	1E-04	3E-04	6E-05	1E-06	3E-04	4E-04	6E-04	5E-04	2E-06	4E-04	3E-04	6E-04	5E-04	0.002	1E-05	4E-05	7E-05	2E-04	0.001	2E-04	1E-05	1E-05	0.002	5E-04	5E-04	5E-04	9E-04	8E-05	2E-04	5E-05	5E-04	5E-04	0.001	6E-05	4E-04	1E-05	8E-05	2E-07	**0.00037**
Bike	6E-05	7E-06	2E-04	2E-04	2E-04	4E-04	6E-06	1E-04	2E-04	2E-06	7E-05	4E-06	1E-04	8E-05	3E-04	2E-04	1E-04	1E-04	3E-04	1E-06	1E-04	2E-05	3E-04	9E-06	2E-06	2E-06	9E-05	1E-04	1E-04	0.001	2E-05	3E-04	4E-06	6E-04	0.002	6E-04	7E-04	1E-04	0.001	**0.00025**
PTcom	4E-04	1E-06	9E-04	8E-04	1E-04	1E-04	0.002	0.001	2E-04	3E-04	1E-04	0.002	0.001	0.001	8E-04	2E-04	0.001	3E-05	6E-05	5E-04	0.001	0.003	9E-04	0.003	0.002	0.003	0.005	0.006	7E-04	0.012	5E-04	3E-04	8E-04	0.006	0.002	3E-04	6E-04	5E-04	0.002	**0.00161**
Walk	5E-04	1E-04	1E-04	3E-04	2E-04	7E-04	0.004	0.004	0.001	4E-04	5E-04	0.005	0.004	0.005	1E-04	6E-04	0.001	1E-04	6E-04	8E-05	0.002	0.002	1E-04	0.001	0.009	0.009	0.011	0.011	0.005	0.003	0.002	0.002	0.006	0.004	3E-04	2E-04	8E-05	5E-06	2E-04	**0.00249**
AVG	**0.002**	**0.001**	**9E-04**	**0.002**	**0.001**	**2E-04**	**0.011**	**0.009**	**0.004**	**5E-04**	**5E-04**	**0.008**	**0.009**	**0.007**	**9E-04**	**0.001**	**0.004**	**8E-04**	**0.001**	**0.001**	**0.001**	**0.010**	**9E-04**	**0.003**	**0.006**	**0.006**	**0.010**	**0.011**	**0.004**	**0.017**	**0.009**	**0.003**	**0.009**	**0.006**	**0.002**	**3E-04**	**7E-04**	**4E-04**	**0.002**	0.00425

Average coefficients of determination in rows Bus, Train, Driver and CarPass is: **0.00565**

E.1.6 Table of coefficients of determination – data set with ONLY ORIGIN-DESTINATION PAIRS USED IN TYPE C MODEL of school commute

A-S-F ŠTBK	x100	x110	x120	x130	x140	x200	x210	x220	x230	x240	x250	x260	x270	x280	x300	x310	x320	x330	x340	x350	x360	x370	x400	x410	x420	x430	x500	x510	x600	x610	x700	x710	x720	x730	x800	x810	x820	x905	x915	AVERAGE
Bus	0.013	0.003	3E-04	7E-04	0.001	0.099	0.104	0.156	0.094	0.002	0.031	0.101	0.147	0.139	0.027	0.006	0.016	0.068	0.004	0.010	0.027	0.022	0.002	0.018	0.031	0.031	6E-07	0.003	0	0.034	5E-04	9E-04	0.005	0.004	0.024	0.007	0.003	0.013	0.031	**0.03282**
Train	0.011	0.008	1E-04	0.011	0.016	0.022	0.023	0.035	0.050	0.008	0.010	0.008	0.029	0.017	0.046	0.009	0.010	8E-04	0.010	0.012	0.001	0.006	0.041	0.015	0.004	0.004	6E-04	3E-06	0	0.129	0.005	0.005	4E-04	0.012	0.078	0.061	0.019	0.088	0.021	**0.02120**
Driver	0.033	2E-04	0.133	0.071	0.005	0.003	0.013	0.014	0.076	0.111	0.030	0.223	0.016	0.119	0.041	0.115	0.050	0.042	0.141	0.530	0.005	0.006	0.122	0.013	0.004	0.004	0.013	0.011	0	0.025	3E-04	8E-04	1E-04	0.008	0.021	0.012	0.093	0.018	0.011	**0.05476**
CarPass	0.027	0.012	0.013	0.001	0.002	0.006	1E-04	0.001	4E-04	0.020	0.038	4E-04	0.001	0.002	3E-04	0.002	0.050	0.150	0.002	0.002	7E-04	4E-04	0.020	0.061	0.006	0.005	0.007	0.004	0	0.005	0.005	0.005	2E-04	0.061	0.010	0.054	0.014	0.021	5E-05	**0.01565**
AVERAGE	**0.021**	**0.006**	**0.037**	**0.021**	**0.006**	**0.033**	**0.035**	**0.052**	**0.055**	**0.035**	**0.027**	**0.083**	**0.048**	**0.069**	**0.029**	**0.033**	**0.032**	**0.065**	**0.039**	**0.138**	**0.008**	**0.009**	**0.046**	**0.027**	**0.011**	**0.011**	**0.005**	**0.005**	**0**	**0.048**	**0.003**	**0.003**	**0.001**	**0.021**	**0.033**	**0.034**	**0.032**	**0.035**	**0.016**	0.03111

E.2 Type B models of school commute

E.2.1 Table of correlation coefficients

B-S-F	x300	x310	x320	x330	x340	x350	x360	x370	x400	x410	x420	x430	x500	x510	x600	x610	x700	x710	x720	x730	x800	x810	x820	x905	x915
Bus	0.06	-0.04	0.09	0.08	0.04	-0.05	-0.01	-0.14	-0.05	-0.05	-0.10	-0.10	-0.12	-0.09	-0.08	0.11	-0.15	-0.06	-0.11	0.05	0.07	0.02	-0.03	0.02	-0.06
Train	0.01	0.01	0.02	-0.06	-0.01	-0.02	0.06	0.07	0.07	-0.09	0.11	0.11	0.10	0.11	0.09	-0.15	0.05	0.07	0.11	-0.11	-0.05	0.01	-0.06	-0.02	0.06
MCT	0.00	0.01	-0.10	-0.06	-0.01	0.05	-0.03	0.14	0.00	0.02	0.07	0.06	0.11	0.10	0.02	-0.01	0.17	0.04	0.07	-0.07	-0.08	0.00	0.03	-0.05	0.07
Driver	-0.02	0.00	0.00	-0.03	0.00	0.01	0.06	0.02	0.04	0.04	0.02	0.02	0.02	0.00	0.02	-0.01	0.02	0.01	0.02	0.01	0.00	-0.01	0.00	0.00	0.00
CarPass	-0.08	0.03	-0.09	0.01	-0.03	0.10	-0.09	0.02	-0.06	0.16	-0.09	-0.09	-0.06	-0.09	-0.06	0.06	0.05	-0.03	-0.04	0.11	0.02	-0.03	0.08	0.02	-0.02
PTplus	-0.04	0.02	-0.01	-0.03	-0.02	-0.02	0.05	0.05	0.03	-0.02	0.06	0.07	0.06	0.07	0.08	-0.06	0.05	0.04	0.07	-0.06	-0.01	0.00	-0.04	0.01	0.01
Bike	-0.01	-0.01	0.01	0.02	0.01	-0.01	-0.03	-0.04	-0.03	0.00	-0.04	-0.04	-0.03	-0.04	-0.04	0.00	-0.02	-0.02	-0.04	0.02	-0.03	-0.04	0.00	0.00	0.02
PTcom	0.05	0.04	0.00	-0.08	-0.04	-0.05	0.08	0.09	0.13	-0.10	0.19	0.19	0.16	0.16	0.14	-0.13	0.07	0.06	0.13	-0.15	-0.05	0.02	-0.08	-0.04	0.06
Walk	-0.01	-0.04	-0.01	0.06	0.04	0.01	-0.06	-0.03	-0.05	0.04	-0.07	-0.08	-0.05	-0.07	-0.07	0.04	-0.03	-0.04	-0.06	0.08	0.00	-0.01	0.11	0.03	-0.02

E.2.2 Table of coefficients of determination

B-S-F	x300	x310	x320	x330	x340	x350	x360	x370	x400	x410	x420	x430	x500	x510	x600	x610	x700	x710	x720	x730	x800	x810	x820	x905	x915	AVERAGE
Bus	0.003	0.002	0.009	0.007	0.002	0.002	2E-04	0.021	0.003	0.002	0.009	0.009	0.015	0.009	0.007	0.011	0.024	0.003	0.012	0.002	0.005	2E-04	7E-04	4E-04	0.004	**0.00652**
Train	7E-05	3E-05	3E-04	0.004	1E-04	5E-04	0.004	0.005	0.004	0.008	0.011	0.011	0.011	0.012	0.009	0.021	0.003	0.005	0.011	0.011	0.002	5E-05	0.003	2E-04	0.004	**0.00570**
MCT	2E-05	5E-05	0.009	0.004	3E-05	0.002	9E-04	0.019	3E-06	4E-04	0.005	0.004	0.012	0.009	6E-04	1E-04	0.028	0.002	0.005	0.004	0.006	3E-06	0.001	0.003	0.005	**0.00486**
Driver	3E-04	3E-06	2E-05	0.001	2E-05	3E-05	0.004	3E-04	0.002	0.002	4E-04	4E-04	2E-04	5E-07	2E-04	2E-04	2E-04	3E-05	3E-04	1E-04	5E-06	5E-05	6E-06	1E-05	2E-05	**0.00045**
CarPass	0.006	0.001	0.009	5E-05	7E-04	0.010	0.008	5E-04	0.004	0.024	0.008	0.008	0.003	0.008	0.004	0.003	0.003	0.001	0.002	0.012	3E-04	8E-04	0.006	2E-04	4E-04	**0.00489**
PTplus	0.002	5E-04	2E-04	0.001	6E-04	3E-04	0.002	0.002	0.001	3E-04	0.004	0.004	0.004	0.006	0.007	0.003	0.002	0.002	0.005	0.004	6E-05	6E-07	0.001	5E-05	3E-05	**0.00209**
Bike	5E-05	2E-04	1E-04	3E-04	2E-04	8E-05	8E-04	0.001	0.001	2E-05	0.002	0.002	9E-04	0.002	0.002	4E-07	6E-04	4E-04	0.002	4E-04	1E-03	0.001	5E-07	6E-06	3E-04	**0.00069**
PTcom	0.003	0.002	6E-07	0.006	0.002	0.003	0.007	0.009	0.016	0.011	0.035	0.035	0.026	0.025	0.020	0.016	0.005	0.004	0.017	0.021	0.003	6E-04	0.006	0.002	0.004	**0.01106**
Walk	5E-05	0.002	4E-05	0.004	0.002	1E-04	0.004	9E-04	0.003	0.001	0.006	0.006	0.002	0.004	0.006	0.002	7E-04	0.002	0.004	0.007	6E-06	5E-05	0.012	8E-04	2E-04	**0.00274**
AVG	**0.002**	**8E-04**	**0.003**	**0.003**	**8E-04**	**0.002**	**0.003**	**0.007**	**0.004**	**0.005**	**0.009**	**0.009**	**0.008**	**0.008**	**0.006**	**0.006**	**0.007**	**0.002**	**0.007**	**0.007**	**0.002**	**3E-04**	**0.003**	**7E-04**	**0.002**	0.00433

E.3 Type C models of school commute

E.3.1 Table of correlation coefficients

C-S-F	x10	x11	x12	x13	x20	x21	x22	x23	x31	x32	x34	x60	x61	x63	x70	x71	x73	x74	x80	x81	x83	x84	x90	x91
Bus	-0.32	0.20	-0.01	-0.10	0.21	-0.17	0.08	-0.36	-0.02	-0.37	0.08	-0.30	-0.28	-0.16	0.69	-0.32	-0.01	-0.16	0.86	-0.39	-0.10	-0.23	0.24	-0.23
Train	0.08	-0.01	0.04	0.20	-0.35	-0.15	-0.37	0.30	-0.07	-0.04	0.00	-0.09	-0.08	-0.13	-0.29	0.72	-0.02	0.07	-0.35	0.97	-0.07	0.01	-0.28	0.17
Driver	0.06	-0.15	-0.11	-0.08	0.18	0.53	0.44	-0.25	-0.04	-0.13	-0.06	-0.11	-0.13	-0.15	-0.11	-0.01	0.27	-0.04	-0.17	-0.04	0.89	-0.10	0.23	-0.20
CarPass	0.10	-0.11	-0.04	-0.15	0.03	-0.09	-0.02	0.15	-0.17	0.04	-0.09	-0.10	-0.09	-0.07	-0.27	-0.04	-0.06	0.50	-0.41	-0.08	-0.09	0.85	0.13	-0.01

E.3.2 Table of coefficients of determination

C-S-F	x10	x11	x12	x13	x20	x21	x22	x23	x31	x32	x34	x60	x61	x63	x70	x71	x73	x74	x80	x81	x83	x84	x90	x91	AVERAGE
Bus	0.102	0.038	1E-04	0.009	0.043	0.030	0.006	0.129	5E-04	0.134	0.006	0.087	0.080	0.027	0.472	0.101	1E-04	0.024	0.741	0.149	0.010	0.055	0.058	0.052	**0.09811**
Train	0.006	1E-04	0.001	0.041	0.120	0.024	0.138	0.088	0.005	0.001	9E-06	0.008	0.006	0.018	0.084	0.516	4E-04	0.005	0.124	0.934	0.004	9E-05	0.080	0.029	**0.09298**
Driver	0.003	0.021	0.011	0.007	0.031	0.281	0.192	0.061	0.002	0.018	0.004	0.012	0.016	0.022	0.012	1E-04	0.070	0.002	0.029	0.002	0.786	0.011	0.051	0.040	**0.07011**
CarPass	0.010	0.012	0.002	0.023	0.001	0.008	5E-04	0.022	0.031	0.002	0.008	0.011	0.009	0.005	0.074	0.002	0.004	0.250	0.165	0.006	0.007	0.724	0.016	4E-05	**0.05801**
AVG	**0.030**	**0.018**	**0.004**	**0.020**	**0.049**	**0.086**	**0.084**	**0.075**	**0.009**	**0.039**	**0.004**	**0.029**	**0.028**	**0.018**	**0.161**	**0.155**	**0.019**	**0.070**	**0.265**	**0.272**	**0.202**	**0.197**	**0.051**	**0.030**	**0.07980**

If omitting the travel time (x70 – x74) and travel time share (x80 – x84) explanatory variables, then the average coefficients of determination in columns x10 – x63, x90 and x91 is: **0.03590**

E.4 Type A models of work commute

E.4.1 Table of correlation coefficients – NONFILTERED data set

A-W-NF	x100	x110	x120	x130	x140	x200	x210	x220	x230	x240	x250	x260	x270	x280	x300	x310	x320	x330	x340	x350	x360	x370	x400	x410	x420	x430	x500	x510	x600	x610	x700	x710	x720	x730	x800	x810	x820	x905	x915
Bus	0.03	-0.03	-0.01	0.03	0.03	0.01	0.06	0.05	-0.04	-0.01	-0.02	-0.04	-0.05	-0.04	0.05	-0.03	0.05	0.02	0.02	-0.03	-0.01	-0.06	0.01	-0.06	-0.01	-0.02	-0.03	-0.01	-0.02	0.08	-0.06	-0.03	-0.04	0.03	0.05	0.04	0.05	0.04	-0.06
Train	0.01	-0.01	0.00	0.01	0.00	0.00	-0.02	-0.02	0.03	0.01	0.01	0.01	0.02	0.01	0.02	-0.01	0.01	-0.01	0.00	-0.04	0.02	0.01	-0.02	-0.07	0.03	0.03	0.03	0.04	0.02	-0.14	0.00	0.01	0.01	-0.05	-0.03	0.00	-0.02	0.00	0.02
MCT	-0.03	0.02	0.02	-0.03	-0.02	0.00	-0.05	-0.05	0.02	0.01	0.01	0.05	0.05	0.04	0.00	0.01	-0.04	-0.01	-0.01	0.00	0.02	0.06	-0.01	-0.02	0.05	0.05	0.06	0.07	0.03	-0.02	0.06	0.03	0.06	-0.04	-0.01	0.02	0.00	-0.01	0.01
Driver	-0.02	0.03	0.00	-0.02	-0.01	-0.01	-0.04	-0.04	0.03	0.00	0.01	0.03	0.04	0.03	-0.06	0.03	-0.04	-0.02	-0.02	0.06	0.00	0.05	0.01	0.11	-0.01	-0.01	-0.01	-0.03	0.00	0.03	0.06	0.04	0.03	0.03	-0.01	-0.03	-0.03	-0.04	0.03
CarPass	0.01	-0.02	0.01	0.00	0.01	0.03	0.03	0.04	-0.03	-0.01	-0.01	-0.03	-0.04	-0.04	0.03	-0.01	0.02	0.00	0.01	-0.03	-0.01	-0.03	0.01	-0.02	0.00	0.00	-0.01	0.00	0.00	0.01	-0.03	-0.03	-0.02	0.00	0.03	0.02	0.00	0.02	-0.03
PTplus	0.00	0.00	0.00	0.00	0.00	-0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	-0.01	0.01	0.02	0.00	-0.02	0.00	-0.02	-0.01	-0.04	0.00	0.00	0.00	0.01	-0.01	-0.02	-0.02	-0.01	-0.02	-0.01	0.01	0.02	0.02	0.03	-0.02
Bike	0.00	0.00	-0.01	0.00	0.00	-0.01	0.03	0.02	-0.01	-0.01	0.00	-0.02	-0.02	-0.02	-0.01	0.00	0.01	0.01	0.00	-0.01	-0.02	-0.03	-0.03	0.00	-0.02	-0.02	-0.02	-0.03	-0.02	-0.02	-0.02	-0.01	-0.03	0.00	-0.06	-0.06	-0.02	-0.02	0.05
PTcom	-0.01	0.00	0.01	-0.01	-0.01	-0.01	-0.03	-0.03	0.02	0.01	0.00	0.04	0.03	0.04	0.01	0.01	-0.01	0.00	-0.01	-0.01	0.00	0.03	-0.02	-0.05	0.03	0.03	0.04	0.06	0.02	-0.04	0.03	0.03	0.03	-0.05	-0.03	0.01	0.01	-0.01	0.03
Walk	0.00	-0.01	0.01	0.00	0.00	0.01	-0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.01	-0.01	0.01	0.01	0.01	0.02	0.01	-0.01	0.00	0.00	0.01	0.00	0.03	0.03	0.01	0.03	-0.03

E.4.2 Table of correlation coefficients

A-W-F	x100	x110	x120	x130	x140	x200	x210	x220	x230	x240	x250	x260	x270	x280	x300	x310	x320	x330	x340	x350	x360	x370	x400	x410	x420	x430	x500	x510	x600	x610	x700	x710	x720	x730	x800	x810	x820	x905	x915
Bus	0.02	-0.03	0.00	0.02	0.01	0.00	0.05	0.04	-0.04	0.00	-0.02	-0.02	-0.04	-0.02	0.06	-0.02	0.04	0.03	0.01	-0.04	-0.02	-0.06	-0.01	-0.11	-0.01	-0.01	-0.01	0.02	-0.02	0.08	-0.06	-0.01	-0.03	-0.01	0.05	0.06	0.05	0.06	-0.05
Train	0.01	-0.01	0.00	0.00	0.00	0.01	-0.03	-0.02	0.03	0.01	0.01	0.00	0.02	0.01	0.02	-0.01	0.01	-0.02	0.00	-0.02	0.02	0.03	-0.02	-0.07	0.02	0.02	0.03	0.04	0.02	-0.16	0.00	0.01	0.02	-0.05	-0.02	0.01	-0.02	0.00	0.01
MCT	-0.05	0.03	0.03	-0.06	-0.03	0.01	-0.08	-0.07	0.04	0.01	0.01	0.08	0.07	0.07	0.01	0.03	-0.07	-0.01	-0.02	0.01	0.03	0.11	-0.02	-0.03	0.09	0.09	0.12	0.12	0.06	-0.04	0.11	0.03	0.09	-0.07	-0.03	0.04	0.01	-0.02	0.03
Driver	0.00	0.01	-0.01	0.00	0.01	0.00	-0.03	-0.03	0.02	0.00	0.01	0.03	0.03	0.02	-0.06	0.01	-0.01	-0.03	0.00	0.08	0.01	0.05	0.05	0.15	-0.01	-0.01	-0.02	-0.06	0.01	0.04	0.06	0.03	0.03	0.07	0.00	-0.04	-0.02	-0.05	0.02
CarPass	0.00	-0.01	0.01	0.00	0.01	0.04	0.03	0.04	-0.03	-0.02	-0.01	-0.04	-0.04	-0.05	0.03	-0.01	0.01	0.00	0.01	-0.04	-0.01	-0.04	0.01	-0.02	0.01	0.01	0.00	0.01	0.01	0.00	-0.03	-0.03	-0.01	-0.01	0.03	0.02	-0.01	0.01	-0.03
PTplus	0.00	0.00	-0.01	0.01	0.00	-0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.01	-0.01	0.01	0.02	0.00	-0.02	0.00	-0.02	-0.01	-0.06	0.00	0.00	0.01	0.03	0.00	-0.02	-0.03	-0.02	-0.01	-0.02	0.02	0.03	0.02	0.04	-0.03
Bike	-0.02	0.02	0.00	-0.02	-0.02	-0.02	0.01	0.00	0.01	0.00	0.00	-0.01	0.00	-0.01	-0.02	0.02	-0.02	0.02	-0.02	-0.02	-0.02	-0.03	-0.06	-0.02	-0.03	-0.03	-0.01	-0.02	-0.04	-0.06	-0.02	0.00	-0.04	-0.04	-0.10	-0.08	-0.04	-0.03	0.09
PTcom	-0.01	0.00	0.01	-0.01	-0.01	0.00	-0.02	-0.02	0.01	0.00	-0.01	0.03	0.02	0.02	0.02	0.00	-0.01	0.00	0.00	-0.01	0.02	0.04	-0.01	-0.05	0.05	0.05	0.04	0.05	0.02	-0.05	0.02	0.02	0.03	-0.04	-0.02	0.04	0.01	0.00	0.02
Walk	-0.01	0.00	0.01	-0.01	0.00	0.00	-0.01	-0.01	0.00	0.01	0.01	0.01	0.01	0.01	0.00	0.01	-0.01	0.00	0.00	0.00	0.02	0.01	0.00	0.00	-0.01	-0.01	0.01	0.01	-0.01	-0.01	0.01	0.00	0.00	0.00	0.03	0.02	0.03	0.03	-0.03

E.4.3 Table of correlation coefficients – data set with ONLY ORIGIN-DESTINATION PAIRS USED IN TYPE C MODEL of work commute

A-W-F ŠTBK	x100	x110	x120	x130	x140	x200	x210	x220	x230	x240	x250	x260	x270	x280	x300	x310	x320	x330	x340	x350	x360	x370	x400	x410	x420	x430	x500	x510	x600	x610	x700	x710	x720	x730	x800	x810	x820	x905	x915
Bus	-0.01	0.00	0.04	-0.03	-0.03	0.09	0.09	0.11	-0.14	-0.08	-0.07	-0.05	-0.11	-0.08	0.25	0.03	-0.03	-0.01	-0.02	0.13	-0.01	-0.11	0.14	-0.03	-0.04	-0.04	-0.11	-0.04	-0.05	0.19	0.01	0.02	-0.02	0.00	0.19	0.13	0.17	0.20	-0.18
Train	-0.14	0.19	-0.02	-0.19	-0.19	0.02	-0.02	-0.01	0.00	-0.06	0.01	0.02	0.01	0.01	-0.02	0.01	-0.14	0.13	-0.01	-0.10	-0.05	0.01	-0.15	-0.09	-0.07	-0.07	-0.01	0.04	-0.03	-0.30	-0.01	-0.03	-0.02	-0.10	-0.01	-0.09	-0.05	-0.07	-0.07
Driver	0.04	-0.03	0.02	0.00	0.00	0.09	0.01	0.04	-0.07	-0.03	-0.01	0.00	-0.04	-0.03	0.06	-0.09	0.09	-0.07	0.08	-0.10	-0.01	0.03	0.07	0.03	-0.04	-0.04	-0.01	0.04	-0.12	0.09	-0.13	-0.14	-0.11	0.05	0.12	0.07	0.01	0.10	-0.10
CarPass	0.01	-0.03	0.00	0.03	0.04	-0.02	-0.02	-0.03	0.04	0.05	-0.04	0.01	0.03	0.01	-0.07	0.10	-0.05	-0.09	-0.10	0.00	-0.03	0.05	0.09	0.01	0.00	0.00	-0.03	-0.01	-0.03	0.04	0.10	0.12	0.05	-0.03	-0.05	-0.05	-0.04	-0.05	0.06
Bike	0.02	-0.08	0.03	0.08	0.11	-0.11	-0.03	-0.06	0.11	-0.01	0.07	0.01	0.06	0.05	-0.17	0.13	-0.04	0.07	-0.14	0.09	-0.05	0.00	-0.15	0.03	-0.07	-0.07	0.03	-0.08	-0.03	-0.06	0.04	0.06	-0.02	0.00	-0.25	-0.18	-0.12	-0.20	0.24

E.4.4 Table of coefficients of determination – NONFILTERED data set

A-W-NF	x100	x110	x120	x130	x140	x200	x210	x220	x230	x240	x250	x260	x270	x280	x300	x310	x320	x330	x340	x350	x360	x370	x400	x410	x420	x430	x500	x510	x600	x610	x700	x710	x720	x730	x800	x810	x820	x905	x915	AVERAGE
Bus	9E-04	0.001	4E-05	1E-03	8E-04	3E-05	0.004	0.003	0.002	7E-05	3E-04	0.002	0.003	0.002	0.003	8E-04	0.002	4E-04	6E-04	7E-04	2E-04	0.004	2E-04	0.003	2E-04	2E-04	8E-04	2E-04	2E-04	0.006	0.004	0.001	0.002	6E-04	0.003	0.001	0.002	0.002	0.003	**0.00152**
Train	8E-05	9E-05	1E-06	4E-05	2E-06	2E-05	6E-04	4E-04	7E-04	2E-04	6E-05	3E-05	4E-04	5E-05	5E-04	4E-05	1E-04	5E-05	4E-06	0.001	4E-04	1E-04	4E-04	0.004	7E-04	7E-04	0.001	0.002	3E-04	0.019	2E-05	1E-04	2E-04	0.003	0.001	5E-06	3E-04	3E-06	5E-04	**0.00097**
MCT	8E-04	4E-04	3E-04	9E-04	4E-04	1E-05	0.003	0.002	5E-04	2E-04	4E-05	0.003	0.002	0.002	2E-05	1E-04	0.001	6E-05	1E-04	2E-05	3E-04	0.004	2E-04	3E-04	0.002	0.002	0.004	0.005	0.001	4E-04	0.004	7E-04	0.003	0.002	9E-05	5E-04	8E-06	7E-05	9E-05	**0.00120**
Driver	3E-04	8E-04	8E-06	4E-04	2E-04	9E-05	0.002	0.001	8E-04	1E-08	1E-04	0.001	0.001	0.001	0.004	7E-04	0.001	5E-04	4E-04	0.004	1E-05	0.003	2E-04	0.011	1E-04	1E-04	6E-05	0.001	5E-06	0.001	0.004	0.001	0.001	0.001	2E-04	9E-04	7E-04	0.002	7E-04	**0.00122**
CarPass	9E-05	3E-04	7E-05	2E-05	1E-04	7E-04	0.001	0.001	8E-04	1E-04	7E-05	0.001	0.001	0.001	9E-04	2E-04	3E-04	6E-06	2E-04	6E-04	9E-05	0.001	2E-04	2E-04	2E-06	9E-07	4E-05	2E-06	1E-05	5E-05	9E-04	0.001	4E-04	2E-07	7E-04	2E-04	5E-06	2E-04	6E-04	**0.00043**
PTplus	5E-07	4E-09	7E-06	1E-05	4E-06	8E-05	6E-05	1E-05	2E-05	2E-05	3E-08	7E-06	1E-05	1E-07	1E-04	4E-05	6E-05	4E-04	1E-05	3E-04	4E-06	3E-04	1E-04	0.002	7E-06	6E-06	6E-06	1E-04	9E-05	3E-04	6E-04	2E-04	3E-04	2E-04	6E-05	3E-04	2E-04	8E-04	3E-04	**0.00018**
Bike	8E-07	9E-06	3E-05	1E-05	4E-06	5E-05	7E-04	4E-04	4E-05	3E-05	1E-06	6E-04	4E-04	4E-04	9E-05	2E-06	3E-05	5E-05	1E-06	2E-04	3E-04	8E-04	6E-04	1E-05	6E-04	6E-04	4E-04	7E-04	4E-04	6E-04	6E-04	2E-04	0.001	1E-05	0.004	0.003	4E-04	3E-04	0.002	**0.00050**
PTcom	1E-04	2E-05	3E-05	6E-05	9E-05	2E-04	0.001	0.001	3E-04	1E-04	3E-06	0.002	0.001	0.001	9E-05	1E-04	1E-04	4E-08	1E-04	2E-04	2E-05	0.001	4E-04	0.003	0.001	0.001	0.002	0.003	3E-04	0.001	7E-04	0.001	0.001	0.002	7E-04	2E-04	3E-05	2E-04	7E-04	**0.00073**
Walk	2E-05	9E-05	7E-05	1E-06	7E-06	1E-04	3E-05	2E-06	9E-06	3E-05	1E-05	6E-06	1E-06	6E-09	1E-04	1E-05	2E-06	9E-06	2E-05	1E-05	4E-04	2E-05	5E-05	2E-04	2E-04	2E-04	1E-04	2E-04	1E-04	5E-05	2E-06	1E-05	3E-05	2E-05	9E-04	9E-04	1E-04	7E-04	9E-04	**0.00015**
AVG	**3E-04**	**3E-04**	**6E-05**	**3E-04**	**2E-04**	**1E-04**	**0.001**	**0.001**	**6E-04**	**8E-05**	**7E-05**	**0.001**	**0.001**	**9E-04**	**9E-04**	**2E-04**	**6E-04**	**2E-04**	**2E-04**	**8E-04**	**2E-04**	**0.002**	**3E-04**	**0.003**	**6E-04**	**6E-04**	**9E-04**	**0.001**	**3E-04**	**0.003**	**0.002**	**6E-04**	**0.001**	**1E-03**	**0.001**	**8E-04**	**4E-04**	**7E-04**	**1E-03**	0.00077

E.4.5 Table of coefficients of determination

A-W-F	x100	x110	x120	x130	x140	x200	x210	x220	x230	x240	x250	x260	x270	x280	x300	x310	x320	x330	x340	x350	x360	x370	x400	x410	x420	x430	x500	x510	x600	x610	x700	x710	x720	x730	x800	x810	x820	x905	x915	AVERAGE
Bus	4E-04	9E-04	2E-06	5E-04	2E-04	2E-05	0.002	0.001	0.001	2E-05	4E-04	6E-04	0.001	6E-04	0.004	3E-04	0.001	7E-04	8E-05	0.002	3E-04	0.003	2E-04	0.011	9E-05	9E-05	2E-04	4E-04	5E-04	0.007	0.003	2E-04	9E-04	5E-05	0.002	0.003	0.002	0.003	0.003	**0.00153**
Train	5E-05	5E-05	2E-05	5E-07	2E-05	2E-04	0.001	6E-04	0.001	2E-04	6E-05	1E-05	5E-04	3E-05	3E-04	6E-05	3E-05	3E-04	5E-06	4E-04	5E-04	8E-04	6E-04	0.005	5E-04	5E-04	0.001	0.002	4E-04	0.025	5E-06	8E-05	5E-04	0.002	4E-04	2E-04	3E-04	1E-05	9E-05	**0.00116**
MCT	0.003	0.001	0.001	0.003	0.001	5E-05	0.007	0.005	0.001	2E-04	1E-04	0.006	0.005	0.004	4E-05	7E-04	0.004	1E-04	6E-04	3E-05	8E-04	0.012	4E-04	9E-04	0.008	0.008	0.013	0.014	0.004	0.001	0.012	0.001	0.008	0.006	9E-04	0.002	2E-04	6E-04	9E-04	**0.00355**
Driver	2E-05	2E-04	2E-04	6E-07	8E-05	2E-05	8E-04	7E-04	3E-04	4E-06	2E-04	6E-04	7E-04	6E-04	0.003	5E-05	2E-04	0.001	1E-06	0.006	1E-04	0.002	0.002	0.022	7E-05	5E-05	4E-04	0.004	2E-04	0.002	0.003	8E-04	9E-04	0.005	1E-05	0.002	4E-04	0.003	3E-04	**0.00161**
CarPass	2E-05	2E-04	1E-04	2E-06	3E-05	0.001	0.001	0.002	0.001	4E-04	1E-04	0.002	0.002	0.002	9E-04	4E-05	2E-04	1E-05	4E-05	0.001	1E-04	0.001	1E-04	3E-04	1E-04	1E-04	1E-05	1E-04	2E-04	2E-05	7E-04	0.001	1E-04	6E-05	1E-03	4E-04	1E-04	2E-04	8E-04	**0.00055**
PTplus	1E-05	4E-06	4E-05	8E-05	2E-05	1E-04	2E-07	1E-05	1E-05	3E-05	9E-06	5E-06	1E-05	2E-05	8E-05	1E-04	1E-04	6E-04	2E-05	4E-04	2E-05	2E-04	2E-04	0.004	8E-06	8E-06	3E-05	7E-04	2E-05	5E-04	8E-04	2E-04	2E-04	4E-04	4E-04	0.001	4E-04	0.002	9E-04	**0.00034**
Bike	5E-04	3E-04	2E-05	2E-04	4E-04	3E-04	2E-04	2E-05	7E-05	2E-07	2E-05	2E-04	9E-06	4E-05	6E-04	3E-04	3E-04	5E-04	3E-04	6E-04	5E-04	6E-04	0.003	4E-04	0.001	0.001	2E-04	3E-04	0.001	0.003	5E-04	1E-06	0.001	0.001	0.011	0.006	0.001	0.001	0.007	**0.00122**
PTcom	6E-05	2E-11	2E-04	1E-04	1E-04	1E-05	4E-04	4E-04	3E-05	2E-05	4E-05	7E-04	3E-04	5E-04	6E-04	3E-07	4E-05	2E-05	1E-05	2E-04	3E-04	0.002	9E-05	0.003	0.002	0.002	0.001	0.002	6E-04	0.003	6E-04	3E-04	9E-04	0.001	4E-04	0.001	8E-05	3E-06	3E-04	**0.00065**
Walk	7E-05	2E-07	1E-04	1E-04	2E-06	2E-05	1E-04	1E-04	6E-06	7E-05	6E-05	2E-04	1E-04	2E-04	2E-05	7E-05	2E-04	4E-06	7E-06	1E-05	3E-04	1E-04	1E-07	2E-06	7E-05	8E-05	3E-05	7E-05	7E-05	6E-05	1E-04	1E-05	2E-07	2E-05	7E-04	4E-04	0.001	9E-04	9E-04	**0.00017**
AVG	**4E-04**	**3E-04**	**2E-04**	**5E-04**	**2E-04**	**2E-04**	**0.001**	**0.001**	**6E-04**	**1E-04**	**1E-04**	**0.001**	**0.001**	**9E-04**	**0.001**	**2E-04**	**8E-04**	**4E-04**	**1E-04**	**0.001**	**3E-04**	**0.003**	**8E-04**	**0.005**	**0.001**	**0.001**	**0.002**	**0.003**	**8E-04**	**0.005**	**0.002**	**4E-04**	**0.001**	**0.002**	**0.002**	**0.002**	**7E-04**	**0.001**	**0.002**	0.00120

Average coefficients of determination in rows Bus, Train, Driver, CarPass and Bike is: **0.00121**

E.4.6 Table of coefficients of determination – data set with ONLY ORIGIN-DESTINATION PAIRS USED IN TYPE C MODEL of work commute

A-W-F STBK	x100	x110	x120	x130	x140	x200	x210	x220	x230	x240	x250	x260	x270	x280	x300	x310	x320	x330	x340	x350	x360	x370	x400	x410	x420	x430	x500	x510	x600	x610	x700	x710	x720	x730	x800	x810	x820	x905	x915	AVERAGE
Bus	6E-05	2E-08	0.002	0.001	6E-04	0.009	0.008	0.011	0.019	0.006	0.006	0.002	0.011	0.006	0.062	1E-03	7E-04	7E-05	5E-04	0.017	4E-05	0.013	0.020	0.001	0.001	0.002	0.012	0.002	0.002	0.036	4E-05	6E-04	3E-04	3E-07	0.036	0.017	0.028	0.039	0.033	**0.01040**
Train	0.020	0.035	6E-04	0.038	0.036	5E-04	4E-04	4E-05	6E-06	0.003	6E-05	5E-04	4E-05	7E-05	3E-04	2E-04	0.018	0.016	1E-04	0.011	0.003	1E-04	0.021	0.009	0.005	0.004	4E-05	0.002	9E-04	0.087	3E-05	0.001	4E-04	0.010	7E-05	0.008	0.003	0.005	0.005	**0.00882**
Driver	0.002	9E-04	5E-04	1E-05	2E-06	0.008	4E-05	0.002	0.005	1E-03	8E-05	4E-06	0.001	7E-04	0.004	0.008	0.007	0.005	0.006	0.009	1E-04	7E-04	0.005	0.001	0.001	0.002	7E-05	0.002	0.014	0.007	0.017	0.020	0.011	0.003	0.015	0.005	1E-04	0.011	0.010	**0.00475**
CarPass	2E-04	7E-04	2E-07	7E-04	0.001	3E-04	6E-04	6E-04	0.002	0.003	0.002	3E-05	7E-04	2E-04	0.005	0.011	0.002	0.009	0.010	7E-07	1E-03	0.002	0.009	1E-04	6E-06	6E-06	1E-03	2E-04	0.001	0.001	0.009	0.015	0.003	0.001	0.002	0.003	0.002	0.002	0.003	**0.00270**
Bike	3E-04	0.006	7E-04	0.007	0.013	0.013	7E-04	0.004	0.011	2E-04	0.005	1E-04	0.004	0.002	0.027	0.017	0.001	0.005	0.019	0.008	0.003	6E-06	0.021	8E-04	0.006	0.005	7E-04	0.007	1E-03	0.004	0.001	0.003	3E-04	5E-07	0.060	0.032	0.014	0.041	0.058	**0.01033**
AVERAGE	**0.004**	**0.009**	**8E-04**	**0.009**	**0.010**	**0.006**	**0.002**	**0.003**	**0.007**	**0.003**	**0.002**	**6E-04**	**0.003**	**0.002**	**0.020**	**0.007**	**0.006**	**0.007**	**0.007**	**0.009**	**0.001**	**0.003**	**0.015**	**0.002**	**0.003**	**0.003**	**0.003**	**0.002**	**0.004**	**0.027**	**0.006**	**0.008**	**0.003**	**0.003**	**0.023**	**0.013**	**0.009**	**0.020**	**0.022**	0.00740

E.5 Type B models of work commute

E.5.1 Table of correlation coefficients

B-W-F	x300	x310	x320	x330	x340	x350	x360	x370	x400	x410	x420	x430	x500	x510	x600	x610	x700	x710	x720	x730	x800	x810	x820	x905	x915
Bus	0.03	-0.01	0.03	0.02	0.00	-0.06	0.06	0.01	0.00	-0.17	0.08	0.08	0.07	0.12	0.06	0.01	0.00	0.05	0.07	-0.09	0.05	0.11	-0.03	0.05	-0.05
Train	0.01	-0.01	0.01	-0.03	0.00	-0.02	0.01	0.04	0.00	-0.07	0.06	0.06	0.07	0.07	0.05	-0.13	0.03	0.04	0.06	-0.05	-0.01	0.02	-0.02	0.01	0.01
MCT	0.01	0.02	-0.06	-0.02	-0.02	0.02	0.01	0.10	-0.01	-0.01	0.06	0.06	0.08	0.08	0.03	-0.03	0.10	0.03	0.07	-0.06	-0.03	0.03	0.03	-0.03	0.03
Driver	-0.03	0.02	-0.04	-0.02	-0.01	0.06	-0.06	0.01	0.02	0.16	-0.05	-0.05	-0.06	-0.09	-0.02	0.06	0.03	-0.01	-0.02	0.09	0.00	-0.06	0.01	-0.06	0.02
CarPass	0.02	-0.01	0.01	0.00	0.01	-0.02	0.03	-0.04	0.02	0.01	-0.02	-0.02	-0.02	-0.03	-0.01	0.02	-0.03	-0.04	-0.03	0.02	0.03	0.01	0.01	0.01	-0.03
PTplus	0.00	-0.01	0.02	0.01	0.01	-0.04	0.05	0.02	-0.01	-0.10	0.08	0.08	0.07	0.10	0.06	-0.06	0.00	0.03	0.05	-0.08	0.02	0.05	-0.02	0.04	-0.03
Bike	-0.01	-0.02	0.02	0.02	0.02	0.00	-0.06	-0.06	-0.04	0.01	-0.09	-0.09	-0.07	-0.08	-0.08	-0.01	-0.06	-0.05	-0.09	0.03	-0.10	-0.10	0.00	-0.02	0.08
PTcom	0.02	0.01	-0.01	-0.02	-0.01	-0.01	0.03	0.05	0.00	-0.05	0.07	0.07	0.07	0.08	0.05	-0.07	0.04	0.04	0.06	-0.07	-0.03	0.03	0.00	-0.01	0.03
Walk	0.00	0.00	0.00	0.00	0.00	0.04	-0.03	0.00	0.01	0.03	-0.04	-0.04	-0.03	-0.03	-0.04	0.02	-0.01	-0.02	-0.03	0.04	0.03	0.02	0.07	0.03	-0.04

E.5.2 Table of coefficients of determination

B-W-F	x300	x310	x320	x330	x340	x350	x360	x370	x400	x410	x420	x430	x500	x510	x600	x610	x700	x710	x720	x730	x800	x810	x820	x905	x915	AVERAGE
Bus	9E-04	7E-05	0.001	6E-04	8E-07	0.003	0.004	2E-04	2E-06	0.029	0.007	0.007	0.005	0.015	0.003	1E-04	2E-05	0.003	0.004	0.009	0.002	0.011	6E-04	0.003	0.002	**0.00445**
Train	7E-05	3E-05	8E-05	0.001	9E-06	5E-04	9E-05	0.002	2E-06	0.005	0.004	0.004	0.004	0.006	0.002	0.018	7E-04	0.002	0.003	0.003	1E-04	6E-04	6E-04	1E-04	5E-05	**0.00222**
MCT	2E-04	4E-04	0.003	4E-04	3E-04	3E-04	1E-04	0.011	7E-05	5E-05	0.003	0.003	0.007	0.006	1E-03	8E-04	0.010	8E-04	0.005	0.003	1E-03	7E-04	0.001	6E-04	9E-04	**0.00241**
Driver	7E-04	4E-04	0.001	4E-04	1E-04	0.004	0.003	6E-05	3E-04	0.025	0.003	0.003	0.003	0.009	4E-04	0.003	8E-04	2E-04	5E-04	0.007	2E-06	0.004	6E-05	0.003	4E-04	**0.00293**
CarPass	2E-04	9E-05	6E-05	2E-06	8E-05	3E-04	0.001	0.002	4E-04	2E-04	3E-04	3E-04	5E-04	9E-04	2E-04	4E-04	9E-04	0.001	8E-04	3E-04	0.001	2E-04	4E-05	2E-04	9E-04	**0.00050**
PTplus	2E-07	2E-04	6E-04	1E-04	5E-05	0.002	0.003	5E-04	8E-05	0.010	0.006	0.006	0.005	0.011	0.003	0.004	2E-05	9E-04	0.003	0.006	4E-04	0.003	5E-04	0.002	8E-04	**0.00270**
Bike	2E-04	3E-04	3E-04	5E-04	3E-04	3E-09	0.004	0.004	0.002	1E-04	0.008	0.008	0.005	0.007	0.007	3E-05	0.003	0.002	0.008	8E-04	0.010	0.010	2E-06	4E-04	0.006	**0.00341**
PTcom	4E-04	1E-04	1E-04	3E-04	1E-04	9E-05	8E-04	0.003	2E-06	0.003	0.006	0.006	0.005	0.006	0.002	0.005	0.002	0.001	0.004	0.004	9E-04	8E-04	4E-06	2E-04	9E-04	**0.00206**
Walk	2E-05	2E-06	2E-05	5E-07	1E-06	0.002	0.001	1E-06	1E-04	0.001	0.002	0.002	9E-04	0.001	0.002	6E-04	6E-05	5E-04	0.001	0.002	0.001	3E-04	0.004	0.001	0.001	**0.00095**
AVG	**3E-04**	**2E-04**	**8E-04**	**4E-04**	**1E-04**	**0.001**	**0.002**	**0.002**	**3E-04**	**0.008**	**0.004**	**0.004**	**0.004**	**0.007**	**0.002**	**0.004**	**0.002**	**0.001**	**0.003**	**0.004**	**0.002**	**0.003**	**8E-04**	**0.001**	**0.001**	**0.00240**

E.6 Type C models of work commute

E.6.1 Table of correlation coefficients

C-W-F	x10	x11	x12	x13	x20	x21	x22	x23	x31	x32	x34	x40	x41	x42	x43	x44	x45	x60	x61	x62	x63	x70	x71	x73	x74	x76	x80	x81	x83	x84	x86	x90	x91
Bus	-0.02	0.06	0.03	0.07	0.20	0.11	0.22	-0.10	0.05	-0.21	0.23	-0.09	-0.04	-0.05	0.04	0.09	0.33	-0.14	-0.14	-0.13	-0.08	0.69	-0.12	-0.20	-0.02	-0.11	0.94	-0.19	-0.30	-0.07	-0.16	0.23	-0.20
Train	0.00	0.02	0.02	-0.04	-0.29	-0.17	-0.34	0.10	-0.05	0.06	-0.06	0.02	-0.01	0.07	0.04	-0.07	-0.11	0.03	0.02	0.02	0.00	-0.16	0.66	-0.13	-0.05	0.06	-0.19	0.94	-0.29	-0.06	-0.08	-0.01	-0.09
Driver	-0.01	-0.04	-0.04	0.02	0.09	0.06	0.11	0.06	0.00	0.10	0.00	-0.01	-0.03	0.08	-0.04	0.01	-0.02	0.07	0.08	0.09	0.10	-0.20	-0.19	0.61	-0.15	-0.14	-0.33	-0.30	0.90	-0.20	-0.34	0.09	-0.10
CarPass	-0.05	0.15	0.09	-0.02	0.03	-0.01	0.02	0.08	-0.03	0.02	0.04	0.16	0.11	0.00	-0.05	-0.02	0.07	-0.02	0.00	-0.03	-0.02	-0.08	-0.03	-0.08	0.70	-0.05	-0.10	-0.07	-0.27	0.89	-0.12	0.02	-0.01
Bike	-0.04	-0.17	-0.17	0.01	-0.08	-0.01	-0.06	-0.08	-0.12	-0.13	-0.17	-0.07	0.04	-0.05	0.02	-0.20	-0.25	-0.18	-0.18	-0.18	-0.17	-0.14	-0.09	-0.34	-0.15	0.35	-0.14	-0.05	-0.33	-0.15	0.93	-0.34	0.36

E.6.2 Table of coefficients of determination

C-W-F	x10	x11	x12	x13	x20	x21	x22	x23	x31	x32	x34	x40	x41	x42	x43	x44	x45	x60	x61	x62	x63	x70	x71	x73	x74	x76	x80	x81	x83	x84	x86	x90	x91	AVERAGE
Bus	6E-04	0.004	0.001	0.005	0.038	0.011	0.050	0.009	0.003	0.044	0.054	0.008	0.002	0.003	0.002	0.007	0.106	0.019	0.019	0.017	0.007	0.479	0.014	0.040	5E-04	0.013	0.879	0.035	0.093	0.006	0.024	0.055	0.039	**0.06323**
Train	5E-06	4E-04	3E-04	0.002	0.082	0.030	0.116	0.011	0.002	0.004	0.004	4E-04	4E-05	0.006	0.001	0.005	0.013	9E-04	4E-04	3E-04	2E-05	0.026	0.442	0.016	0.003	0.004	0.036	0.880	0.086	0.004	0.007	3E-05	0.008	**0.05422**
Driver	4E-05	0.002	0.002	4E-04	0.008	0.003	0.012	0.003	6E-06	0.010	7E-06	3E-05	6E-04	0.006	0.002	1E-04	3E-04	0.005	0.007	0.008	0.010	0.040	0.037	0.378	0.021	0.020	0.106	0.088	0.819	0.041	0.116	0.009	0.011	**0.05351**
CarPass	0.002	0.024	0.008	6E-04	0.001	6E-05	3E-04	0.006	7E-04	3E-04	0.002	0.025	0.013	2E-09	0.003	3E-04	0.005	3E-04	3E-06	0.001	4E-04	0.006	8E-04	0.006	0.485	0.002	0.010	0.006	0.071	0.791	0.013	4E-04	1E-04	**0.04501**
Bike	0.002	0.028	0.027	1E-04	0.006	7E-05	0.004	0.006	0.015	0.017	0.031	0.005	0.002	0.003	5E-04	0.039	0.060	0.034	0.032	0.032	0.029	0.019	0.008	0.114	0.023	0.121	0.019	0.003	0.110	0.024	0.865	0.116	0.128	**0.05827**
AVG	**0.002**	**0.019**	**0.013**	**0.003**	**0.045**	**0.015**	**0.061**	**0.012**	**0.007**	**0.025**	**0.030**	**0.013**	**0.006**	**0.006**	**0.003**	**0.017**	**0.061**	**0.020**	**0.019**	**0.020**	**0.015**	**0.190**	**0.167**	**0.185**	**0.178**	**0.053**	**0.350**	**0.337**	**0.393**	**0.289**	**0.342**	**0.060**	**0.062**	**0.05485**

If omitting the travel time (x70 – x76) and travel time share (x80 – x86) explanatory variables, then the average coefficients of determination in columns x10 – x63, x90 and x91 is: **0.01391**

[F] Computation script of the used decision tree regressor model

In [1]:
```
#1
# import libraries
import pandas as pd
import math
import numpy as np
from statistics import mean, stdev
```

In [2]:
```
#2
# load input data
data = pd.read_csv('B_zam_F4_Bus_v1.csv', sep=';') # filename.csv, column separator: ';'
```

In [3]:
```
#3
# mean and standard deviation of the last column (y variable)
mean(data.Bus.values), stdev(data.Bus.values)
```

Out[3]:
```
(0.12926378783401057, 0.2274481722401821)
```

In [4]:
```
#4
# import libraries
from sklearn.tree import DecisionTreeRegressor
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error
%matplotlib inline
import matplotlib.pyplot as plt
from sklearn.model_selection import cross_val_score
```

In [4]:

```
#4
# import libraries
from sklearn.tree import DecisionTreeRegressor
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error
%matplotlib inline
import matplotlib.pyplot as plt
from sklearn.model_selection import cross_val_score
```

In [5]:

```
#5
# split the dataset into training set and test set
# X - test set with x variables
# y - training set with y variable
# Xtest - test set with x variables
# ytest - test set with y variable
# Function train_test_split shuffles the dataset and divides it into the Xtrain, Xtest, ytrain and ytest
# The input data for the 'x variables' are all the columns excep 'OBECOP', 'OBECPS', 'Index' and 'Bus'
# The input data for 'y variable' is the coulm 'Bus'
# The percentual split is 20% test data and 80% train data
# Random seed is set to allow for reproducibility

Xtrain, Xtest, ytrain, ytest = train_test_split(data.drop(['OBECOP', 'OBECPS', 'Index', 'Bus' ],
                                                axis = 1), data.Bus, test_size=0.2, random_state=42)
# show number of rows and columns in each resulting data set
print(Xtrain.shape)
print(Xtest.shape)
print(ytrain.shape)
print(ytest.shape)
```

```
(29924, 25)
(7482, 25)
(29924,)
(7482,)
```

Baseline Tree Regressor with default parameters

Dokumentation: https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html?highlight=decision%20tree%20regressor#sklearn.tree.DecisionTreeRegressor

In [6]:

```
#6
dtr = DecisionTreeRegressor()    # dtr - Decision Tree Regressor object (model)
# default values of hyperparameters of the model are used (see bellow, or reffer to documentation)
# max_depth: int, default=None
# min_samples_leaf: int or float, default=1
# min_samples_split: int or float, default=2
dtr.fit(Xtrain,ytrain) # fit the model to the train data (i.e. train the model)

# use the trained model to predict y's from Xtrain (dtr.predict(Xtrain))
# and compute the RMSE (root mean squared error) between the predicted y's and the actual train set y's (ytrain)
print('Root mean squared error (train):', \
      np.sqrt(mean_squared_error(dtr.predict(Xtrain), ytrain)))

# use the trained model to predict y's from the previously unseen test data Xtest (dtr.predict(Xtest))
# and compute the RMSE between the predicted ŷ's and the actual test y's (compute the generalization error)
print('Root mean squared error (test):', \
      np.sqrt(mean_squared_error(dtr.predict(Xtest), ytest)))
```

```
Root mean squared error (train): 0.1883510367310464
Root mean squared error (test): 0.23971880501150455
```

Finding better parameters

In [7]:

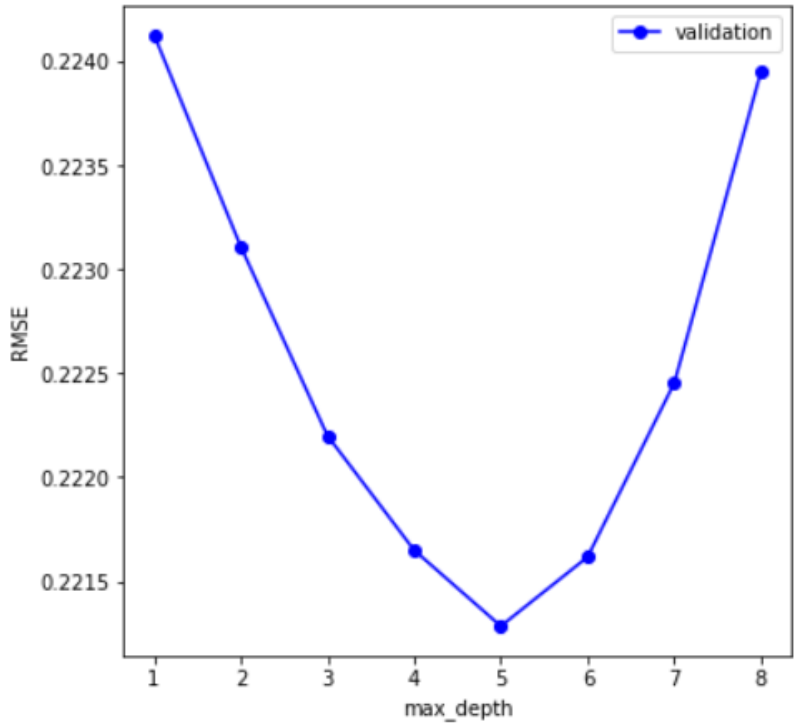
```
#7
# Search for the optimal depth of decision tree
depths = range(1,9,1) # try values from 1 to 8, increment by 1 (1,2,3,4,5,6,7,8)
folds = 5 # number of folds in cross-validation (cv)
rmse = []

# construct a different tree with each depth
for depth in depths:
    dtr = DecisionTreeRegressor(max_depth=depth) # max_depth = hyperparameter of the model
    # for each tree, 5 error values are computed from the 5-fold cross validation
    # use train data, the error is computed as negative mean squared error (NMSE)
    dtr_scores = cross_val_score(dtr, Xtrain, ytrain, cv=folds, scoring='neg_mean_squared_error')
    # the absolute value is taken from each of the 5 error values, then the square root is taken from their average
    # to transform the NMSE into RMSE, thus obtaining 1 RMSE for each tree
    rmse.append(np.sqrt(mean(abs(dtr_scores)))) # append the RMSE into a list
```

In [8]:

```
#8
# visualise the RMSE of individual trees
plt.figure(figsize=(6,6)) # set figure size
plt.plot(depths,rmse,'ob-') # x-axis: depth of tree, y-axis: RMSE; o circles, b blue, - connected by a line
plt.xlabel('max_depth') # x-axis label
plt.ylabel('RMSE') # y-axis label
plt.legend(['validation']) # graph legend
```

Out[8]: <matplotlib.legend.Legend at 0xca98ab0>



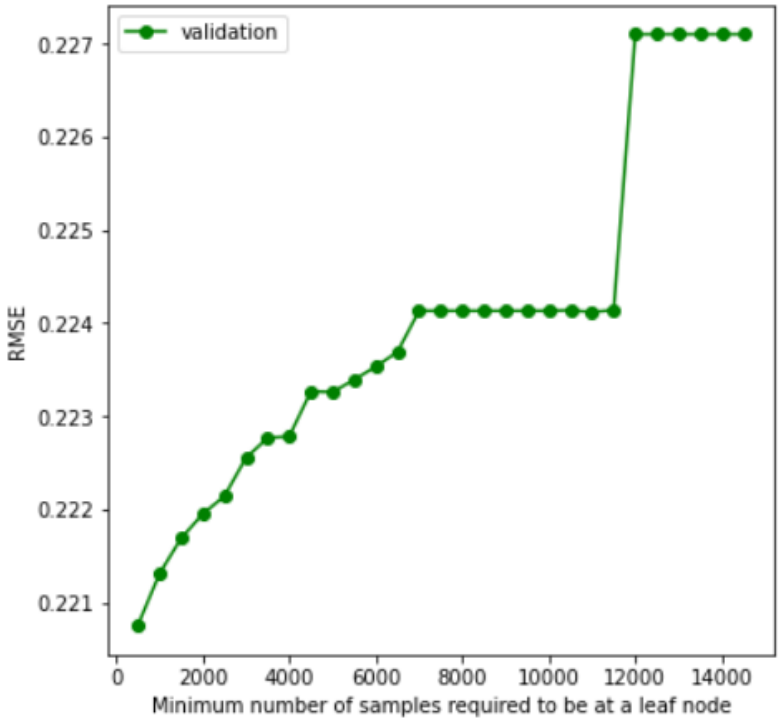
In [9]:

```
#9
# search for the optimal minimal number of data samples in the tree leaf
min_leaf = range(500,15000,500) # from 500 to 14.999, increment by 500
rmse = []
for leaf in min_leaf:
    dtr = DecisionTreeRegressor(min_samples_leaf=leaf)
    dtr_scores = cross_val_score(dtr, Xtrain, ytrain, cv=folds, scoring='neg_mean_squared_error')
    rmse.append(np.sqrt(mean(abs(dtr_scores))))
```

In [10]:

```
#10
plt.figure(figsize=(6,6))
plt.plot(min_leaf,rmse,'og-')
plt.xlabel('Minimum number of samples required to be at a leaf node')
plt.ylabel('RMSE')
plt.legend(['validation'])
```

```
Out[10]: <matplotlib.legend.Legend at 0xcae9ad0>
```



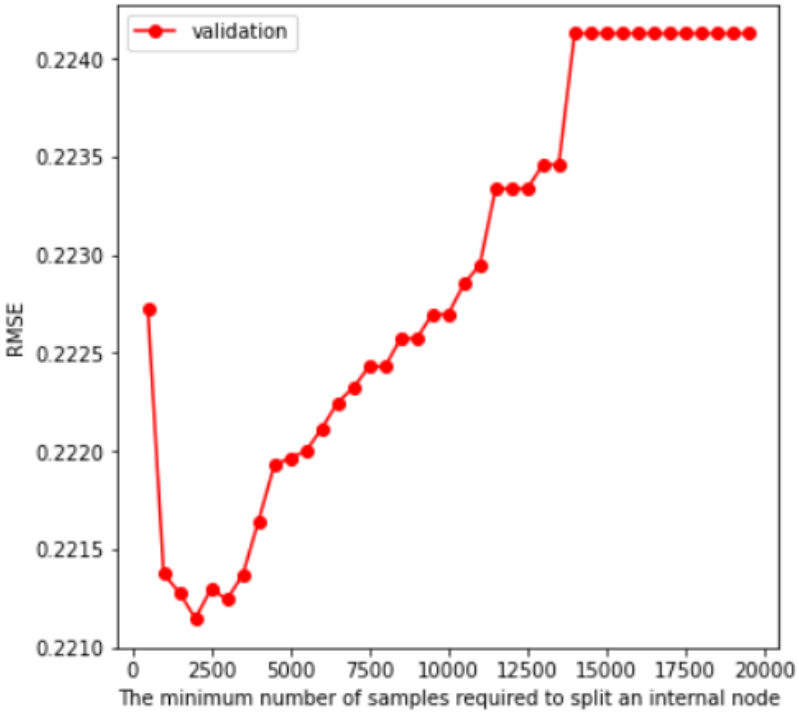
In [11]:

```
#11
# search for the optimal minimal number of data samples requiered to split an internal node
min_samples = range(500,20000,500) # from 500 to 19.999, increment by 500
rmse = []
for sample in min_samples:
    dtr = DecisionTreeRegressor(min_samples_split=sample)
    dtr_scores = cross_val_score(dtr, Xtrain, ytrain, cv=folds, scoring='neg_mean_squared_error')
    rmse.append(np.sqrt(mean(abs(dtr_scores))))
```

In [12]:

```
#12
plt.figure(figsize=(6,6))
plt.plot(min_samples, rmse,'or-')
plt.xlabel('The minimum number of samples required to split an internal node')
plt.ylabel('RMSE')
plt.legend(['validation'])
```

```
Out[12]: <matplotlib.legend.Legend at 0xcb2a9b0>
```



Dokumentation of GridSearchCV

https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html?highlight=gridsearchcv#sklearn.model_selection.GridSearchCV

In [13]:

```
#13
# use "GridSearchCV" to find the best combination of hyperparameters

# import GridSearchCV
from sklearn.model_selection import GridSearchCV

# creation of grid matrix with a selection of values based on previous search
param_grid = {
    'max_depth': [1,2,3,4,5,6],
    'min_samples_leaf': [500,4000,5000,7000,11000],
    'min_samples_split': [2000,3000,5500,8000,9000,10000]
}

# new model
dtr = DecisionTreeRegressor()

# grid search with 5-fold cross-validation, verbose 2 = dont't print the computation information
grid_search = GridSearchCV(estimator = dtr, param_grid = param_grid,
                           cv = 5, n_jobs = -1, verbose = 2)
# n_jobs: number of parallel computations, 1=none, -1=use all cores
```

In [14]:

```
#14
# run grid search on train data
grid_search.fit(Xtrain, ytrain)
# get the best parameters found by grid search
grid_search.best_params_
```

```
Fitting 5 folds for each of 180 candidates, totalling 900 fits
```

```
[Parallel(n_jobs=-1)]: Using backend LokyBackend with 2 concurrent workers.
[Parallel(n_jobs=-1)]: Done  37 tasks      | elapsed:    7.7s
[Parallel(n_jobs=-1)]: Done 158 tasks      | elapsed:   21.3s
[Parallel(n_jobs=-1)]: Done 361 tasks      | elapsed:   50.2s
[Parallel(n_jobs=-1)]: Done 644 tasks      | elapsed:  1.5min
[Parallel(n_jobs=-1)]: Done 900 out of 900 | elapsed:  2.1min finished
```

Out[14]:

```
{'max_depth': 6, 'min_samples_leaf': 500, 'min_samples_split': 2000}
```

In [15]:

```
#15
# evaluation of model with the chosen hyperparameters
dtr = DecisionTreeRegressor(max_depth=6, min_samples_leaf=1, min_samples_split=2) # pseudo-default settings
dtr.fit(Xtrain,ytrain)
print('Parameters:', dtr.get_params())
print('Depth of tree:', dtr.get_depth())
print('Leaves on tree:', dtr.get_n_leaves())
print('Root mean squared error (train):', np.sqrt(mean_squared_error(dtr.predict(Xtrain), ytrain)))
print('Root mean squared error  (test):', np.sqrt(mean_squared_error(dtr.predict(Xtest), ytest)))
```

```
Parameters: {'ccp_alpha': 0.0, 'criterion': 'mse', 'max_depth': 6, 'max_features': None, 'max_leaf_nodes': None, 'min
_impurity_decrease': 0.0, 'min_impurity_split': None, 'min_samples_leaf': 1, 'min_samples_split': 2, 'min_weight_frac
tion_leaf': 0.0, 'presort': 'deprecated', 'random_state': None, 'splitter': 'best'}
Depth of tree: 6
Leaves on tree: 56
Root mean squared error (train): 0.21785836789462548
Root mean squared error  (test): 0.22340220501616917
```

In [16]:

```
#16
# standard deviation of the y values
stdev(data.Bus.values)
```

Out[16]: 0.2274481722401821

In [17]:

```
#17
# names of cloumns in RMSE table
RMSE_columns=['RMSE_train','RMSE_test']

# prepare the table with those 2 columns and 2 rows named: Standard_deviation, Tree_regressor
RMSE_table = pd.DataFrame(data = None, columns=RMSE_columns, \
                    index=['Standard_deviation','Tree_regressor'])

# save the standard deviation into the table
RMSE_table.loc['Standard_deviation'] = stdev(data.Bus.values)

# save both RMSE of the Tree regressor into the table
RMSE_table.loc['Tree_regressor'] = (np.sqrt(mean_squared_error(dtr.predict(Xtrain), ytrain)),\
                                    np.sqrt(mean_squared_error(dtr.predict(Xtest), ytest)))
```

In [18]:

```
#18
# print the table
RMSE_table
```

Out[18]:

	RMSE_train	RMSE_test
Standard_deviation	0.227448	0.227448
Tree_regressor	0.217858	0.223402

In [19]:

```
#19
# preparation of table with the same columns as Xtrain data with empty row 'Tree_regressor'
# for the results of this method
feature_importances = pd.DataFrame(data = None, columns=Xtrain.columns, \
                                   index=['Tree_regressor'])

# dtr (Tree Regressor model) is currently trained on my data and this line of code
# extracts the feature importances, i.e. how much the individual X variables contributed
# to the decision about node splitting
# save feature importances into table
feature_importances.loc['Tree_regressor'] = dtr.feature_importances_
```

In [20]:

```
#20
# print table
feature_importances
```

Out[20]:

	x300	x310	x320	x330	x340	x350	x360	x370	x400	x410	...	x610	x700	x7
Tree_regressor	0.00308055	0.0339914	0.00222853	0.00852005	0.00395875	0.0325198	0.0356561	0.00334418	0	0.468381	...	0.0689322	0.0213222	0.0662

1 rows × 25 columns

Dokumentation of decision tree diagram export: https://scikit-learn.org/stable/modules/tree.html#tree

Dokumentation of graphviz: https://scikit-learn.org/stable/modules/generated/sklearn.tree.export_graphviz.html#sklearn.tree.export_graphviz

In [21]:

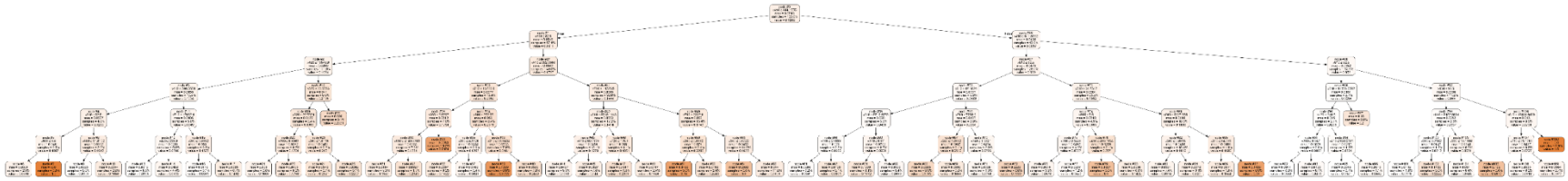
```
#21
# import libraries
from sklearn import tree
import graphviz
from graphviz import Source
```

In [22]:

```
#22
# use Source from graphviz to export and save the decision tree (DT)
# tree.export_graphviz creates the graphical representation of DT
exp_dtr=Source(tree.export_graphviz(dtr, feature_names=Xtrain.columns, filled=True, \
        node_ids=True, proportion=True, rounded=True, special_characters=True, precision=4)\
                                                        ,filename="Bus_dtr", format="png")

exp_dtr # show what is saved for export
```

Out[22]:



In [23]:

```
#23
exp_dtr.view()  # export tree into .png file
```

Out[23]: 'Bus_dtr.png'

In [24]:

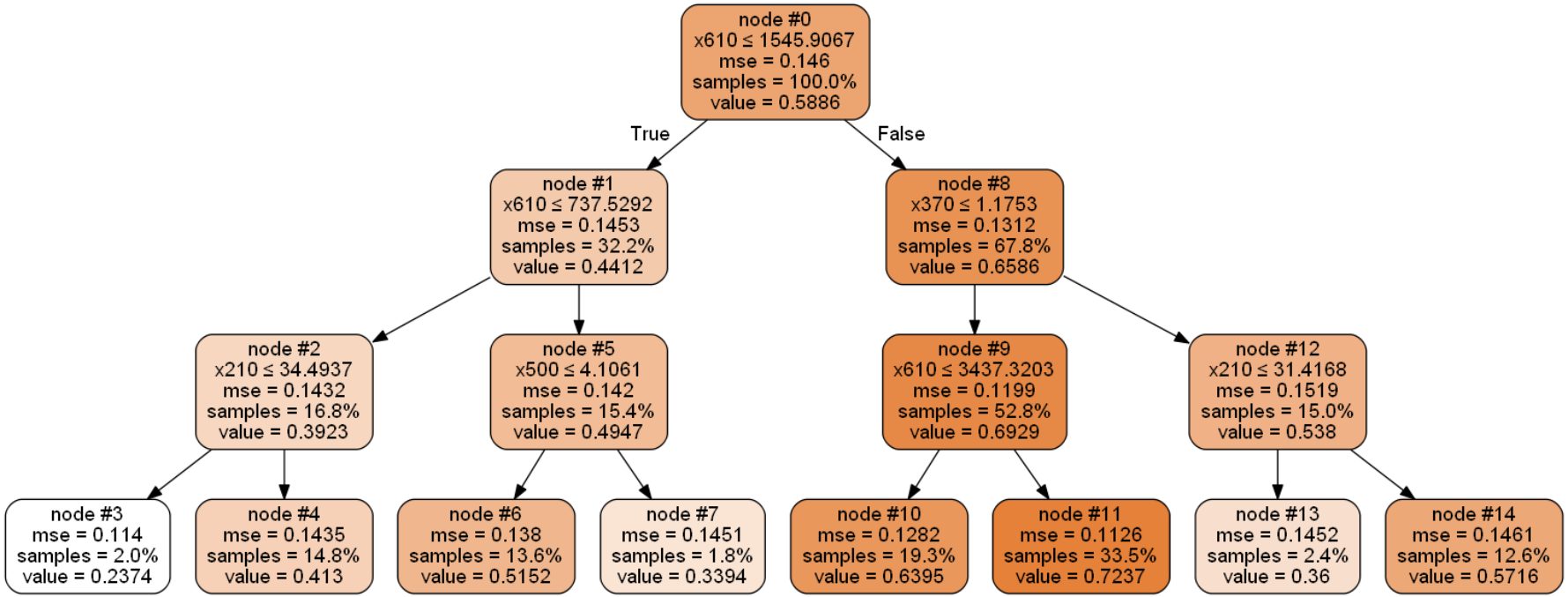
```
#24
RMSE_table.to_csv('Bus_RMSE_table.csv') # save table into RMSE.csv
feature_importances.to_csv('Bus_feature_importances.csv') # save table feature_importances
```

[G] Decision tree diagrams

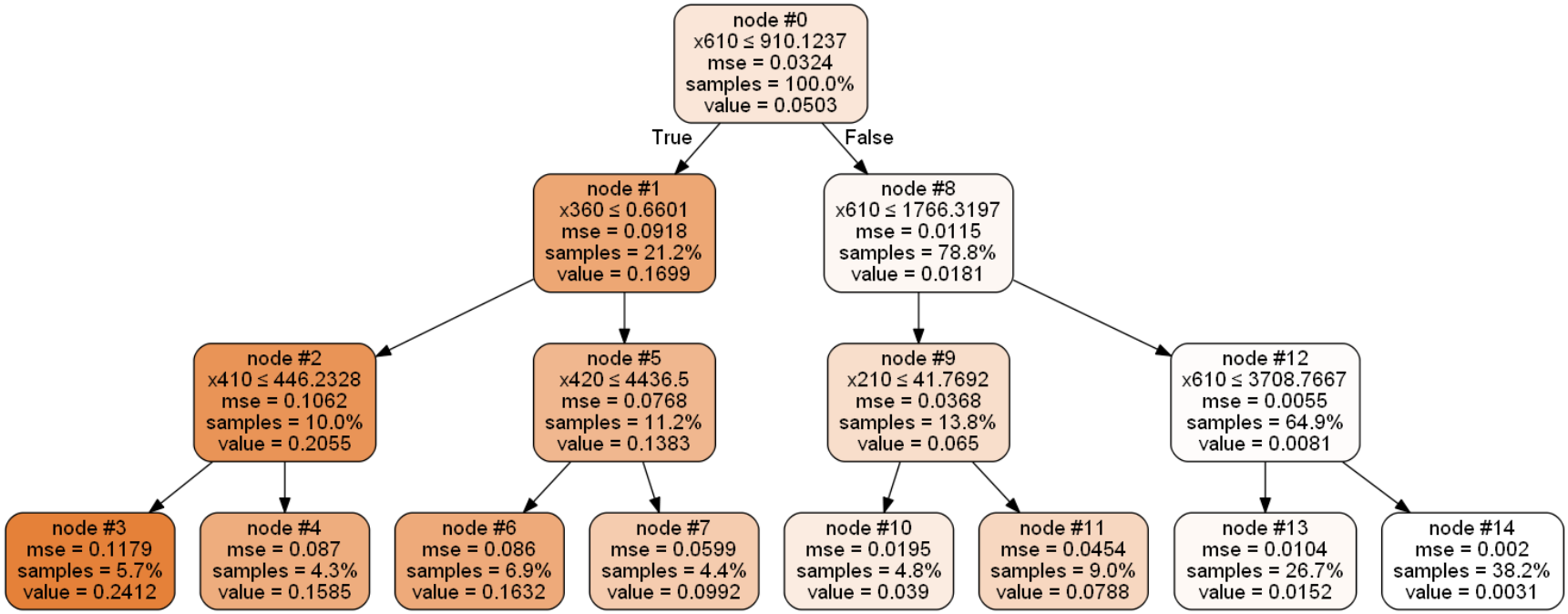
G.1 Type A models of school commute

Decision tree diagrams from type A models of school commute, **version 3**, using 17 explanatory variables and filtered data set of origin-destination pairs.

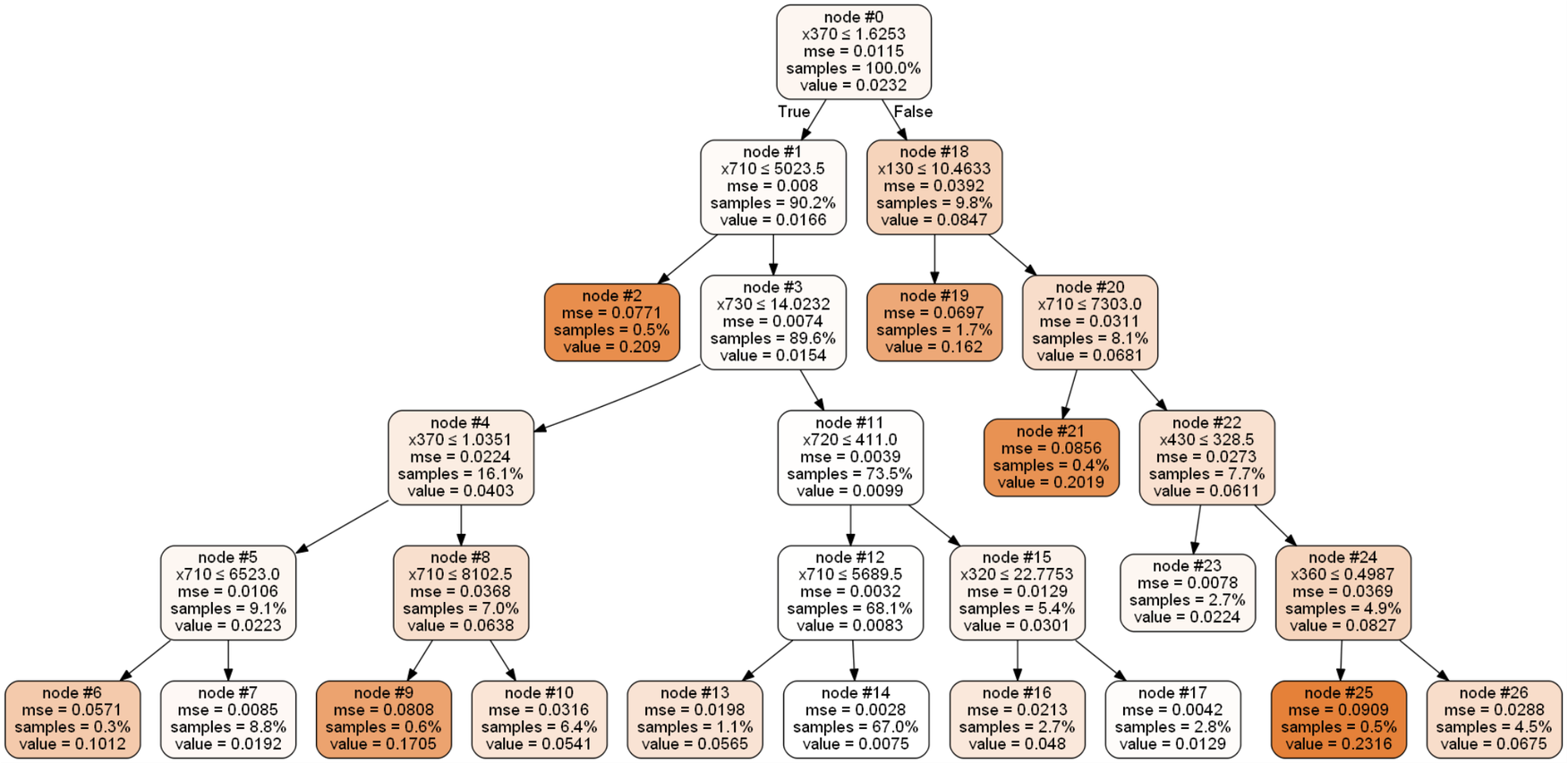
G.1.1 Decision tree diagram of transport mode Bus



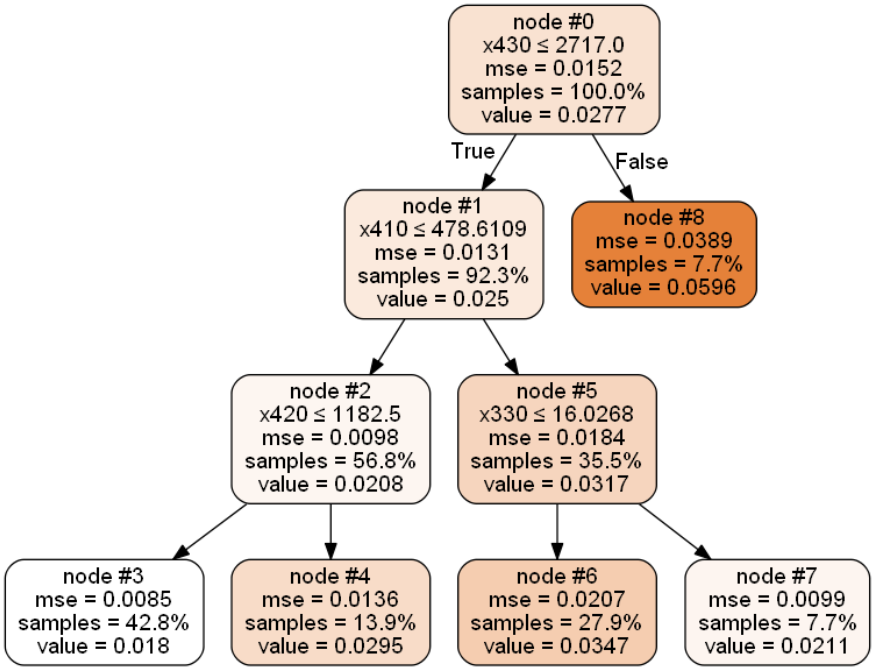
G.1.2 Decision tree diagram of transport mode Train



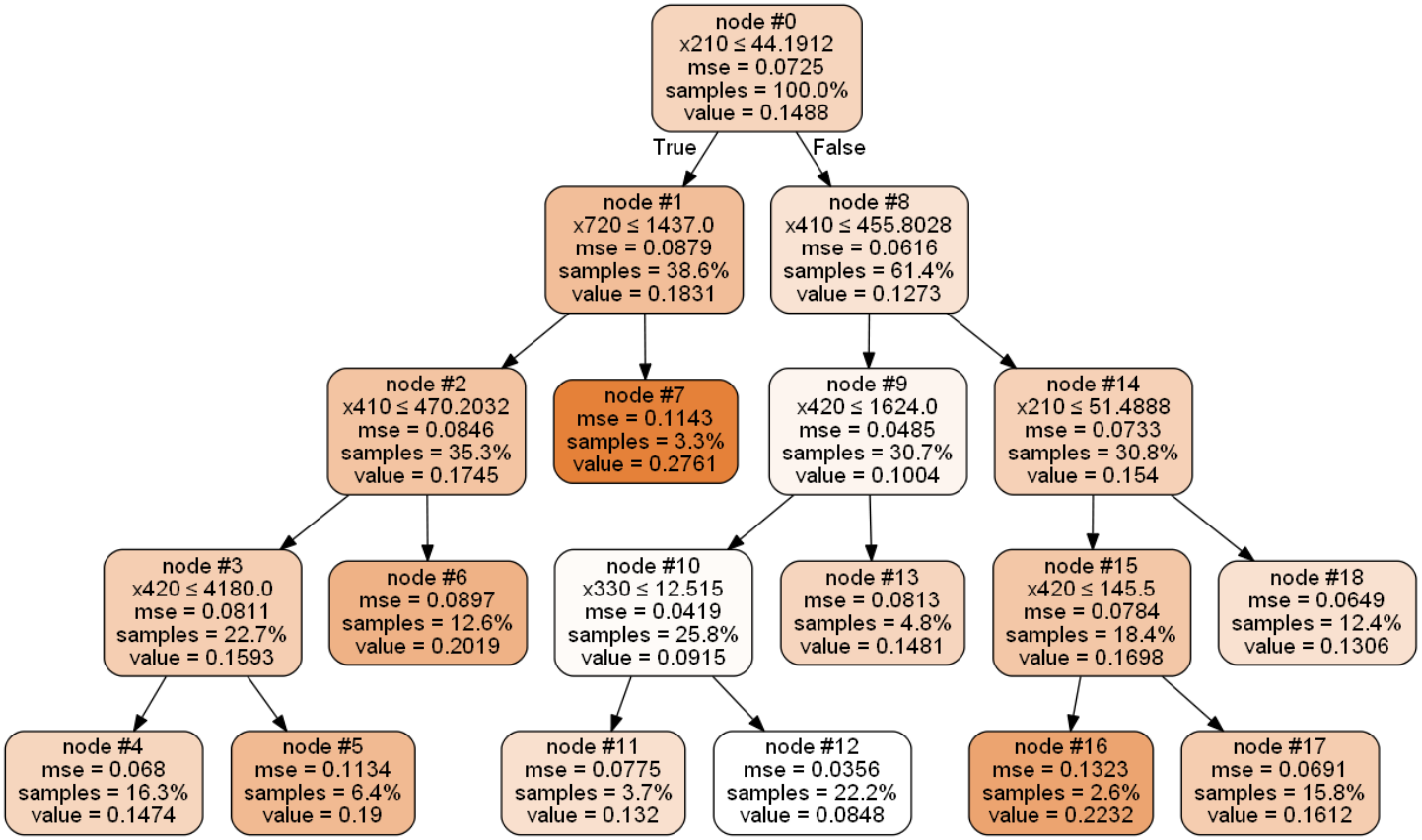
G.1.3 Decision tree diagram of transport mode MCT



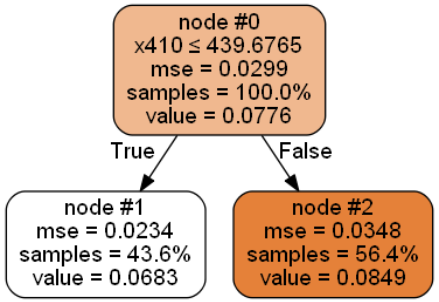
G.1.4 Decision tree diagram of transport mode Driver



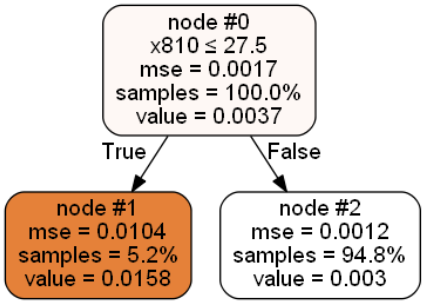
G.1.5 Decision tree diagram of transport mode CarPass



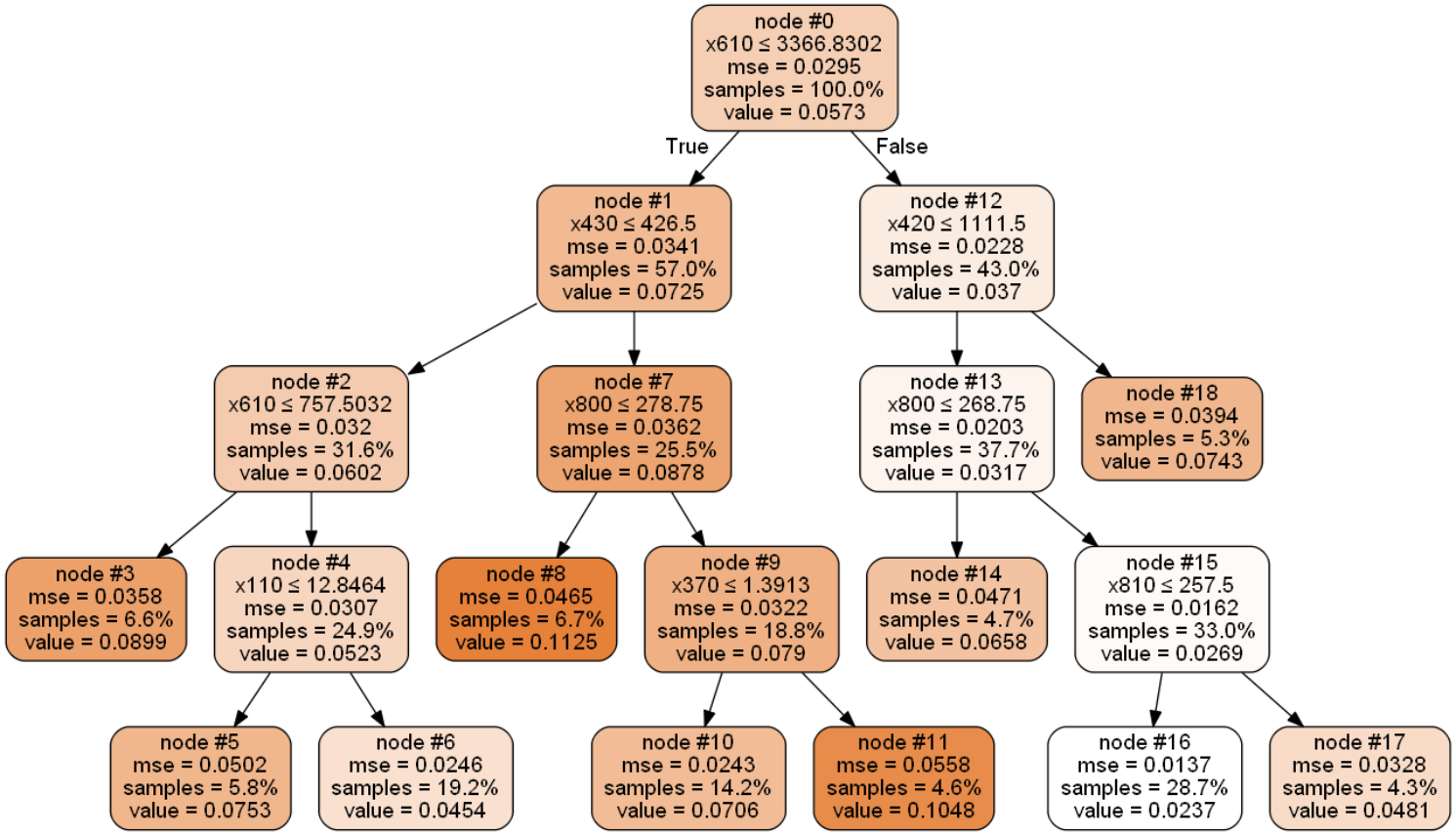
G.1.6 Decision tree diagram of transport mode PTplus



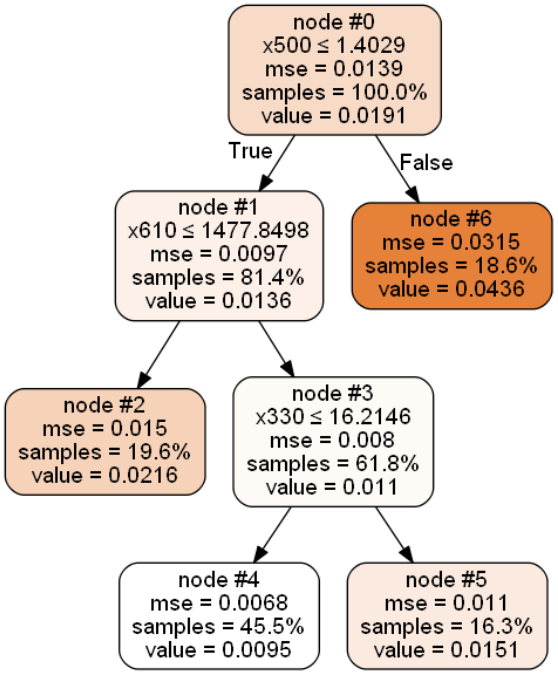
G.1.7 Decision tree diagram of transport mode Bike



G.1.8 Decision tree diagram of transport mode PTcom

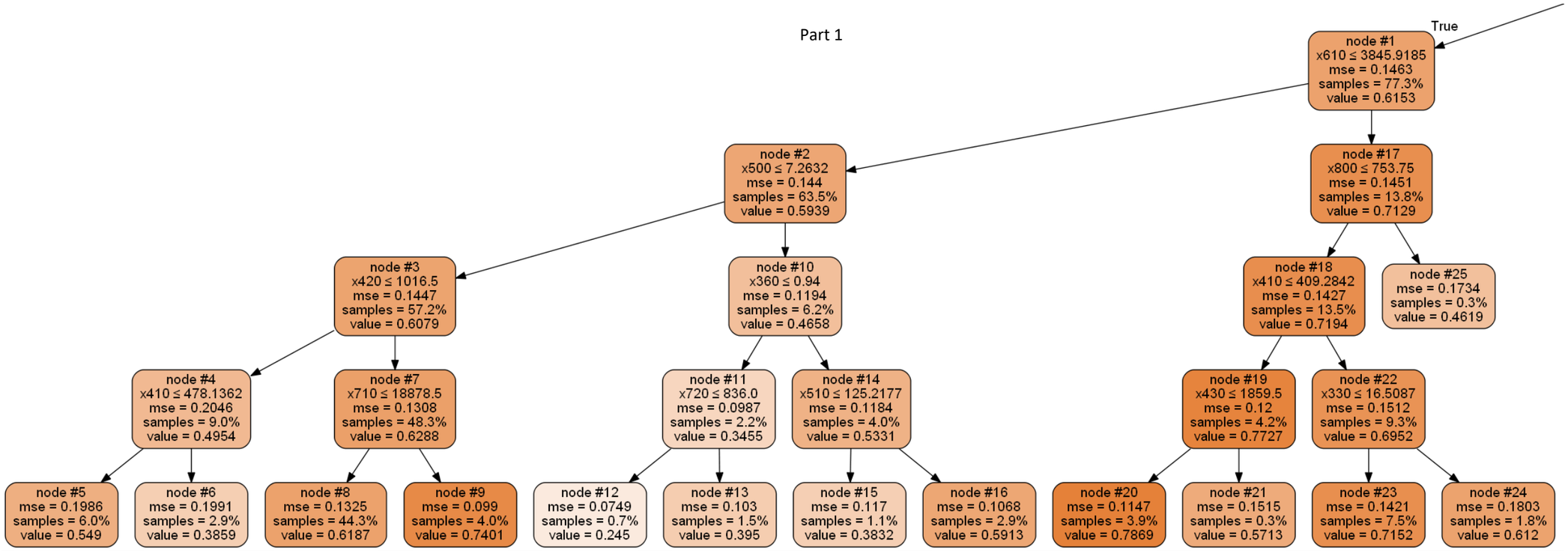


G.1.9 Decision tree diagram of transport mode Walk

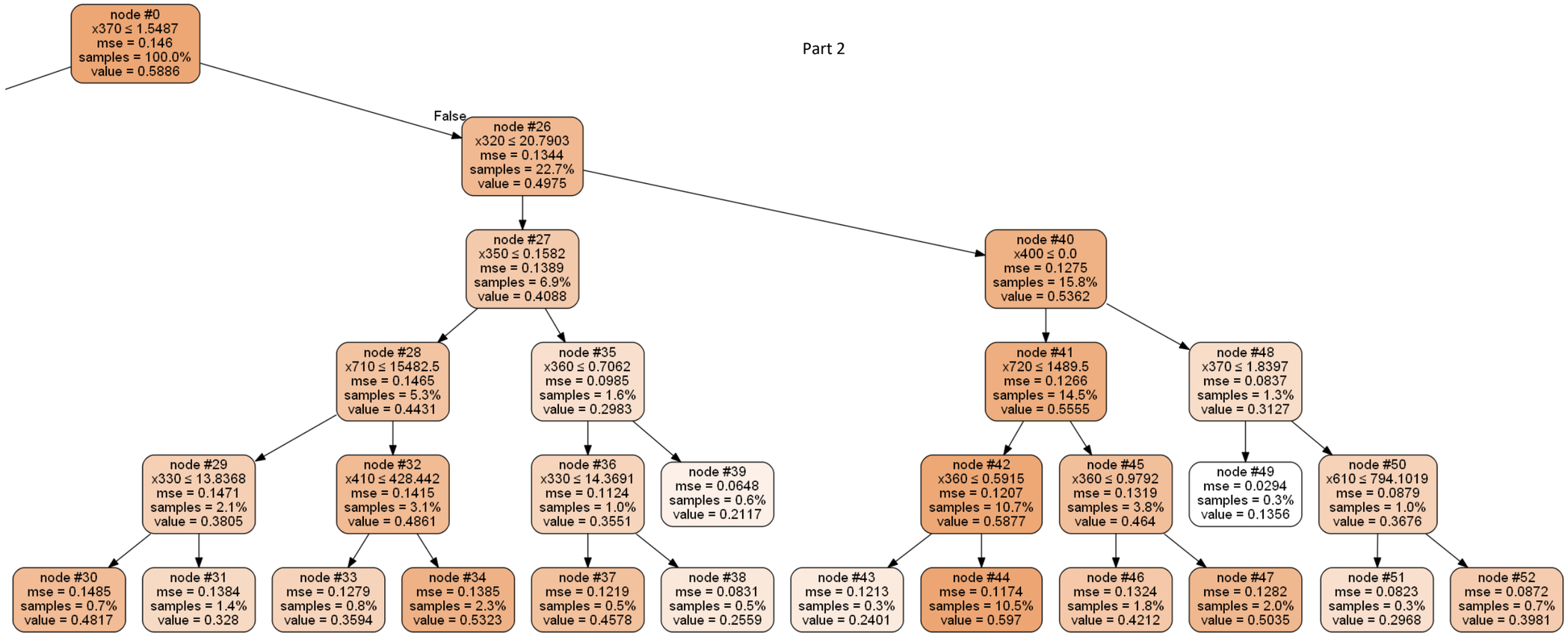


G.2 Type B models of school commute

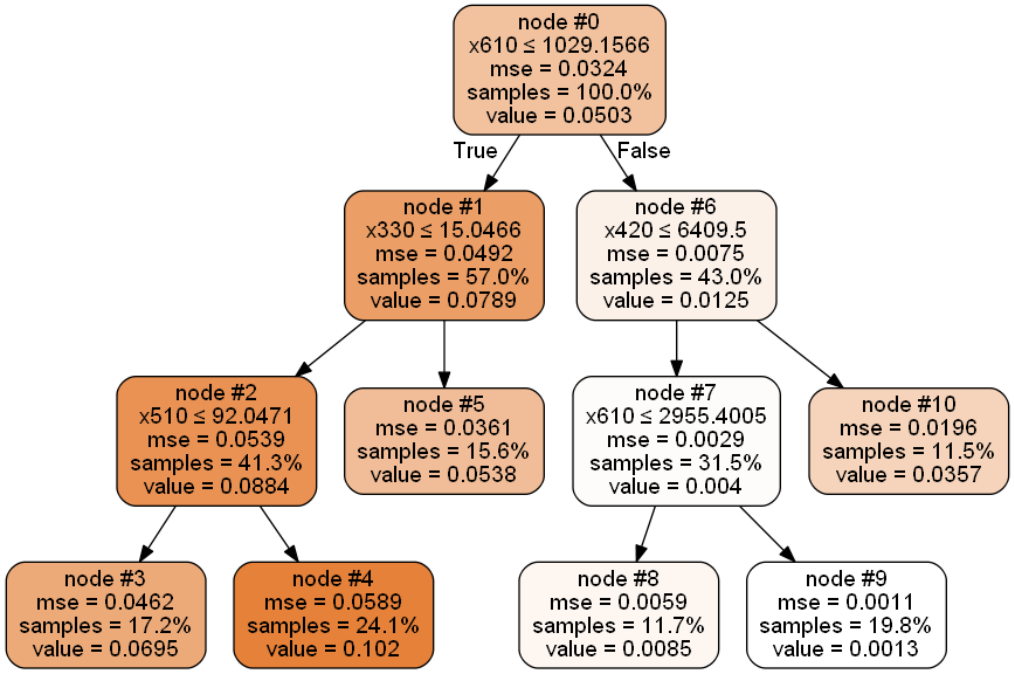
G.2.1 Decision tree diagram of transport mode Bus



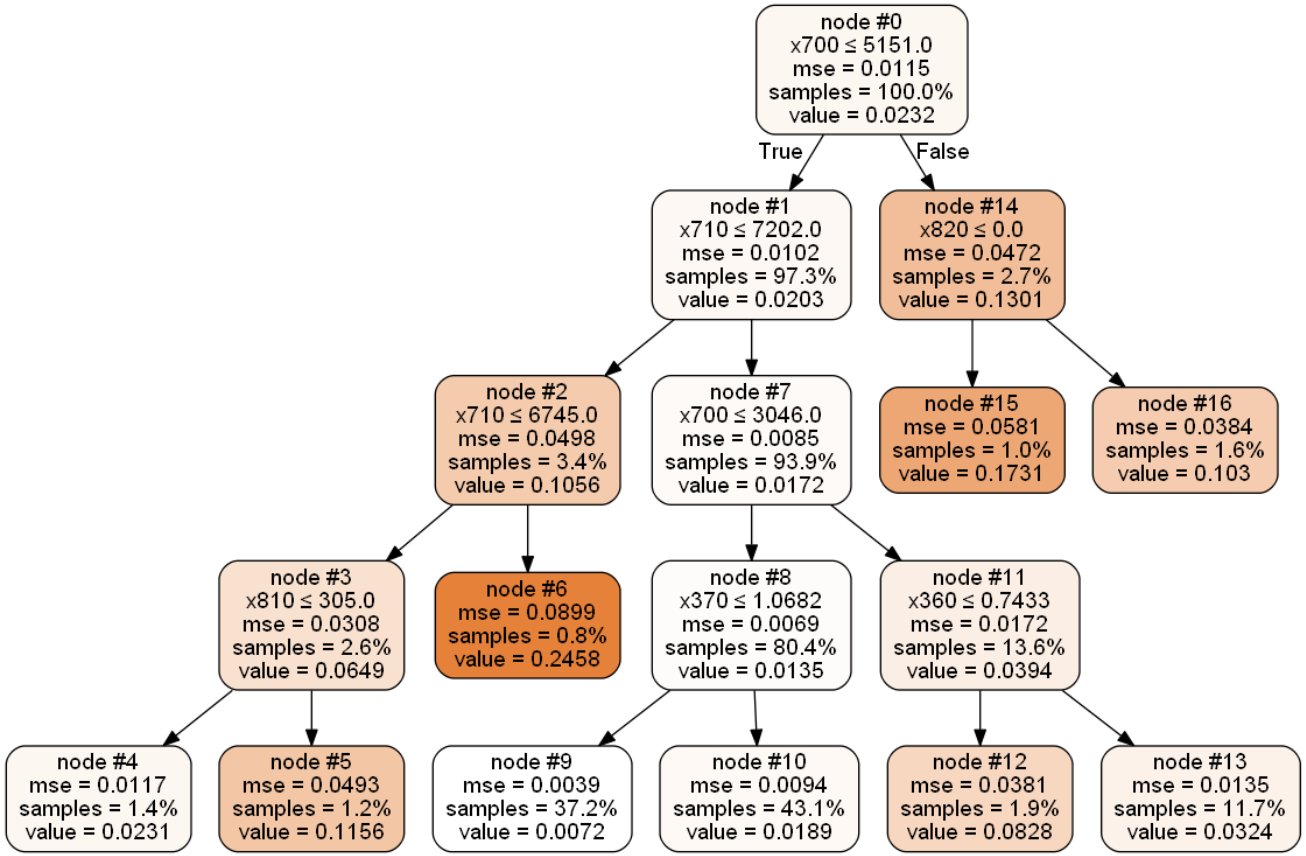
Part 2


node #0
x370 ≤ 1.5487
mse = 0.146
samples = 100.0%
value = 0.5886
False
node #26
x320 ≤ 20.7903
mse = 0.1344
samples = 22.7%
value = 0.4975
node #27
x350 ≤ 0.1582
mse = 0.1389
samples = 6.9%
value = 0.4088
node #40
x400 ≤ 0.0
mse = 0.1275
samples = 15.8%
value = 0.5362
node #28
x710 ≤ 15482.5
mse = 0.1465
samples = 5.3%
value = 0.4431
node #35
x360 ≤ 0.7062
mse = 0.0985
samples = 1.6%
value = 0.2983
node #41
x720 ≤ 1489.5
mse = 0.1266
samples = 14.5%
value = 0.5555
node #48
x370 ≤ 1.8397
mse = 0.0837
samples = 1.3%
value = 0.3127
node #29
x330 ≤ 13.8368
mse = 0.1471
samples = 2.1%
value = 0.3805
node #32
x410 ≤ 428.442
mse = 0.1415
samples = 3.1%
value = 0.4861
node #36
x330 ≤ 14.3691
mse = 0.1124
samples = 1.0%
value = 0.3551
node #39
mse = 0.0648
samples = 0.6%
value = 0.2117
node #42
x360 ≤ 0.5915
mse = 0.1207
samples = 10.7%
value = 0.5877
node #45
x360 ≤ 0.9792
mse = 0.1319
samples = 3.8%
value = 0.464
node #49
mse = 0.0294
samples = 0.3%
value = 0.1356
node #50
x610 ≤ 794.1019
mse = 0.0879
samples = 1.0%
value = 0.3676
node #30
mse = 0.1485
samples = 0.7%
value = 0.4817
node #31
mse = 0.1384
samples = 1.4%
value = 0.328
node #33
mse = 0.1279
samples = 0.8%
value = 0.3594
node #34
mse = 0.1385
samples = 2.3%
value = 0.5323
node #37
mse = 0.1219
samples = 0.5%
value = 0.4578
node #38
mse = 0.0831
samples = 0.5%
value = 0.2559
node #43
mse = 0.1213
samples = 0.3%
value = 0.2401
node #44
mse = 0.1174
samples = 10.5%
value = 0.597
node #46
mse = 0.1324
samples = 1.8%
value = 0.4212
node #47
mse = 0.1282
samples = 2.0%
value = 0.5035
node #51
mse = 0.0823
samples = 0.3%
value = 0.2968
node #52
mse = 0.0872
samples = 0.7%
value = 0.3981

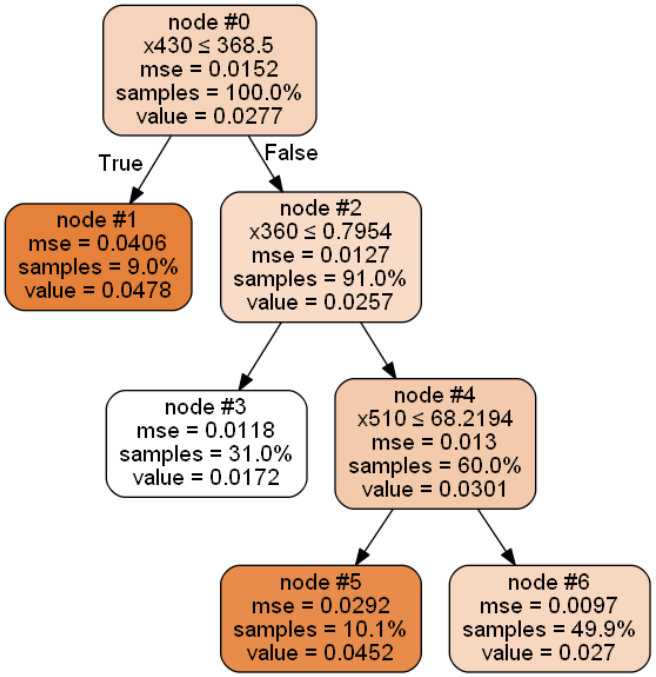
G.2.2 Decision tree diagram of transport mode Train



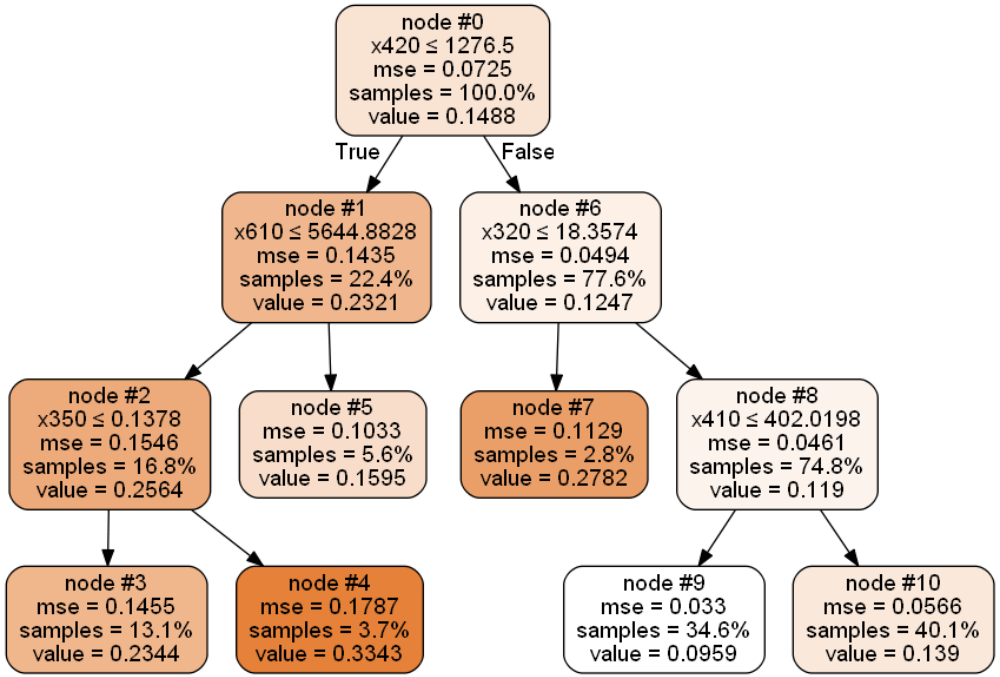
G.2.3 Decision tree diagram of transport mode MCT



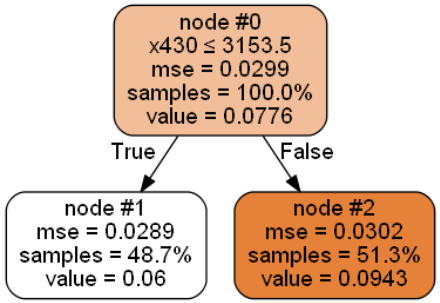
G.2.4 Decision tree diagram of transport mode Driver



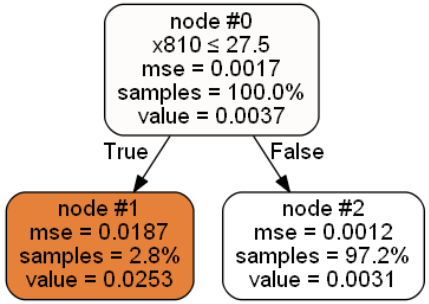
G.2.5 Decision tree diagram of transport mode CarPass



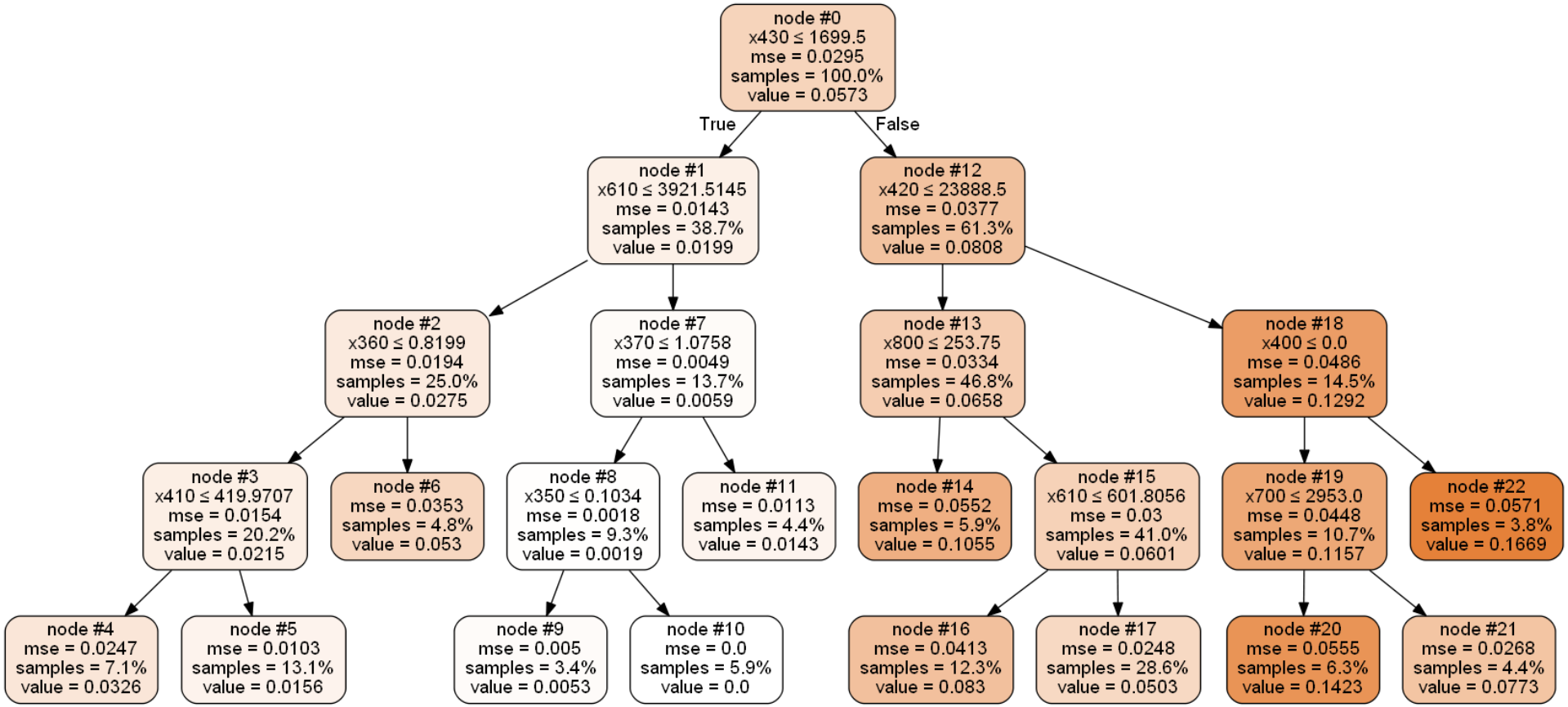
G.2.6 Decision tree diagram of transport mode PTplus



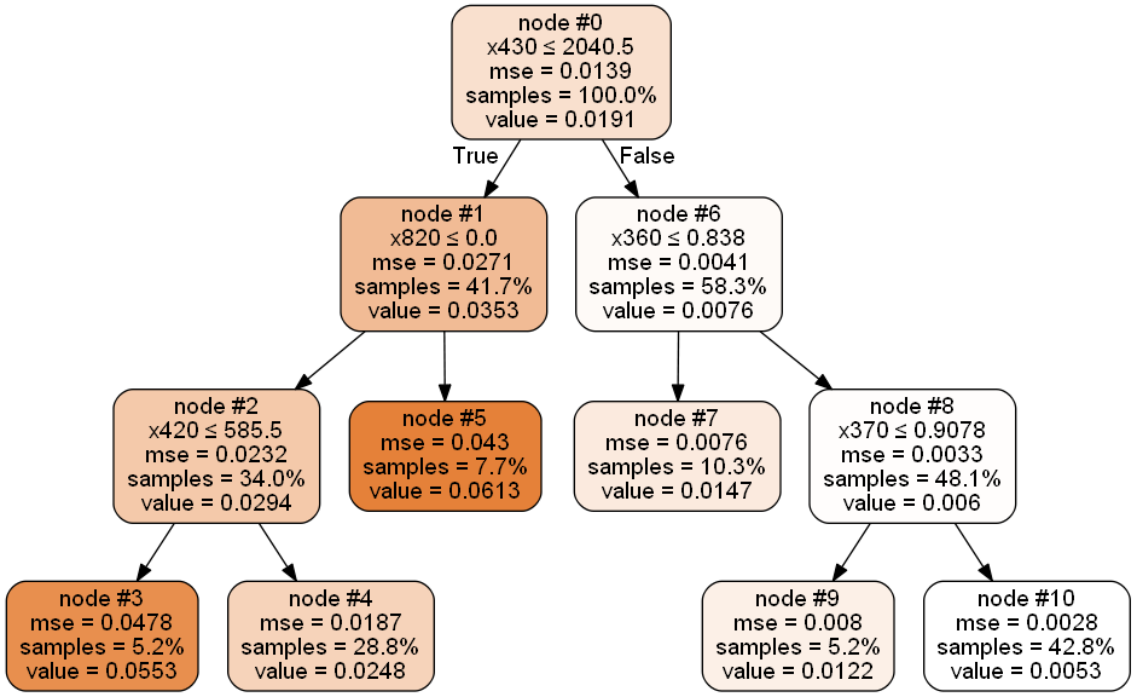
G.2.7 Decision tree diagram of transport mode Bike



G.2.8 Decision tree diagram of transport mode PTcom

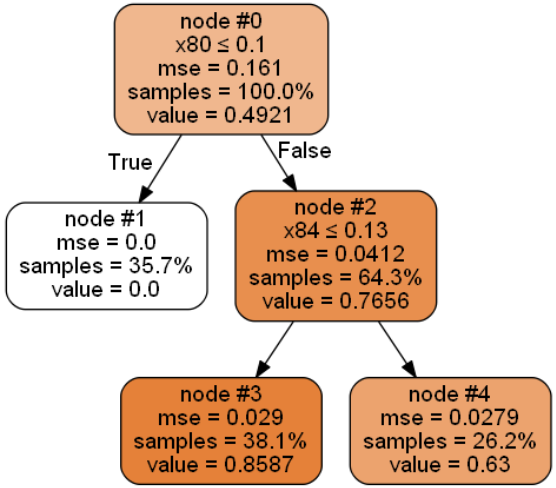


G.2.9 Decision tree diagram of transport mode Walk

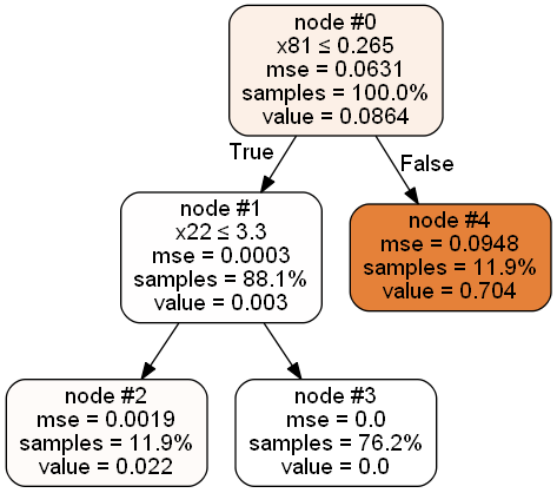


G.3 Type C models of school commute

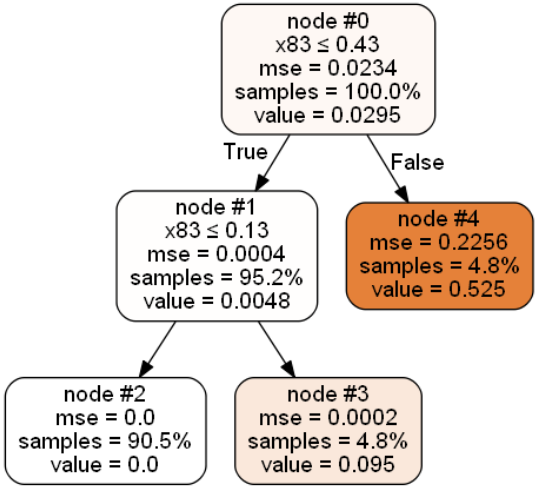
G.3.1 Decision tree diagram of transport mode Bus



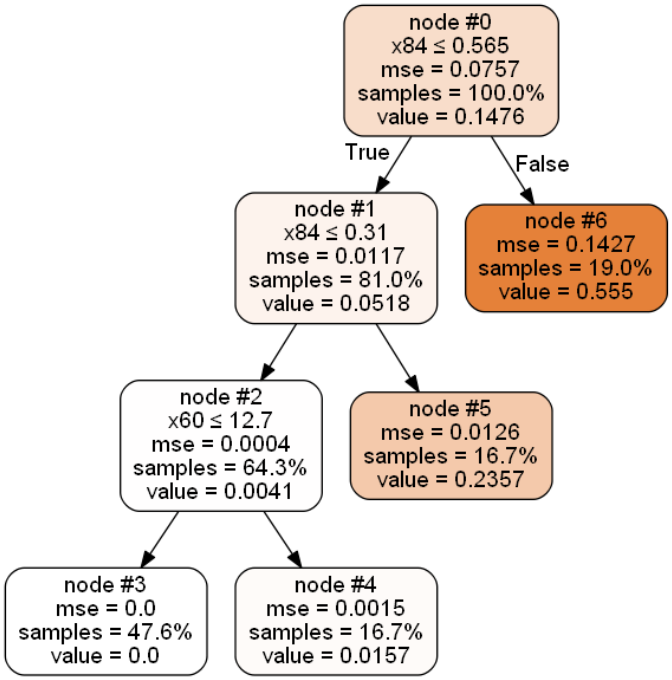
G.3.2 Decision tree diagram of transport mode Train



G.3.3 Decision tree diagram of transport mode Driver

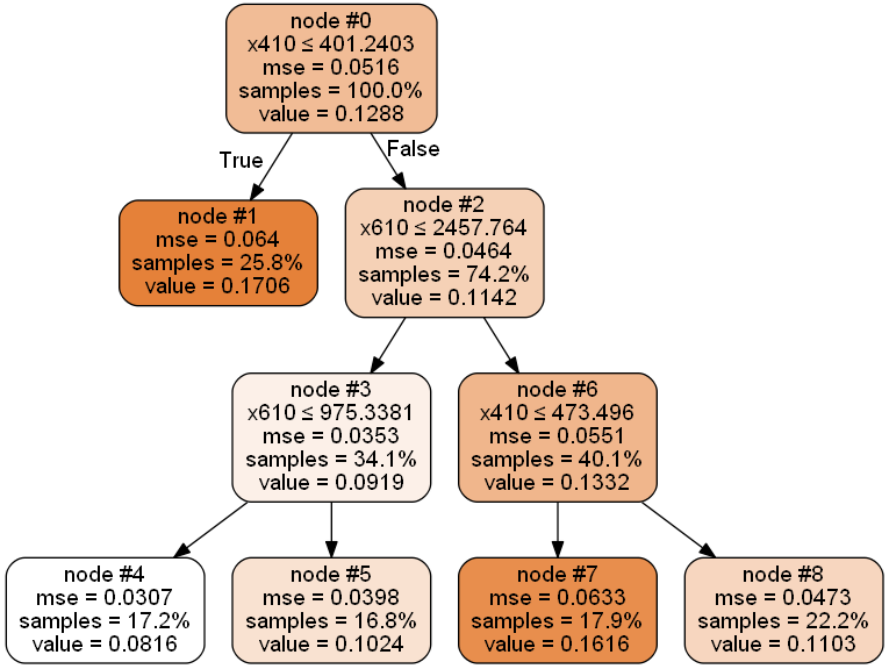


G.3.4 Decision tree diagram of transport mode CarPass

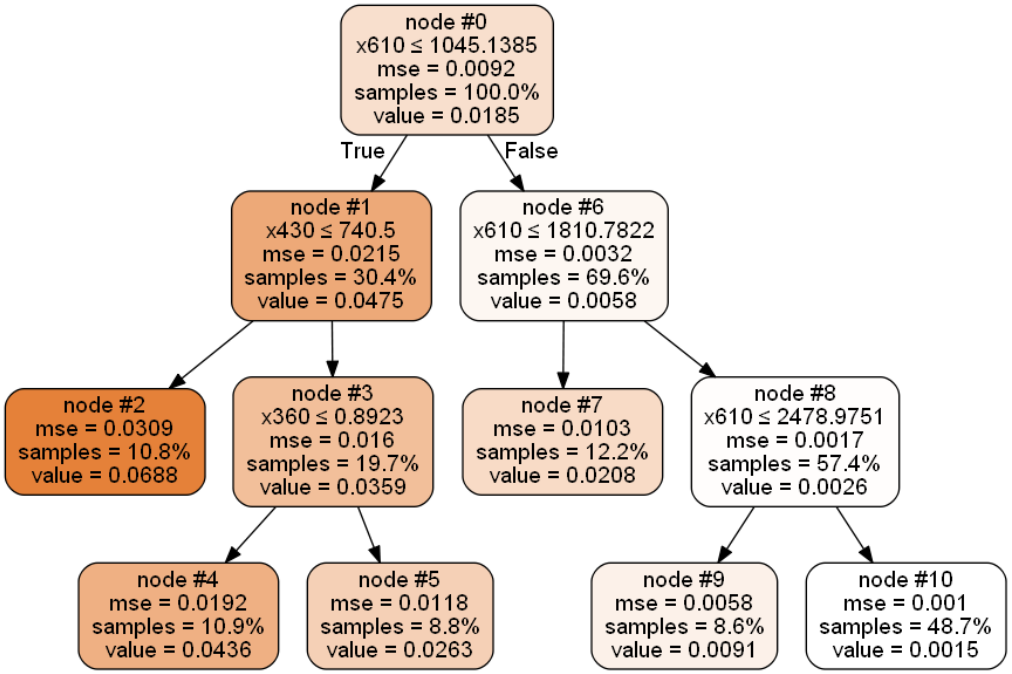


G.4 Type A models of work commute

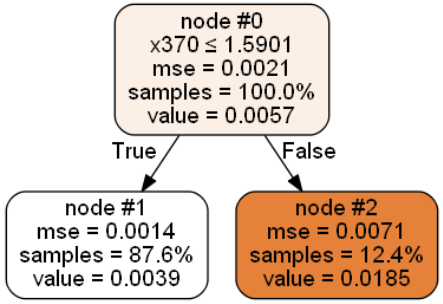
G.4.1 Decision tree diagram of transport mode Bus



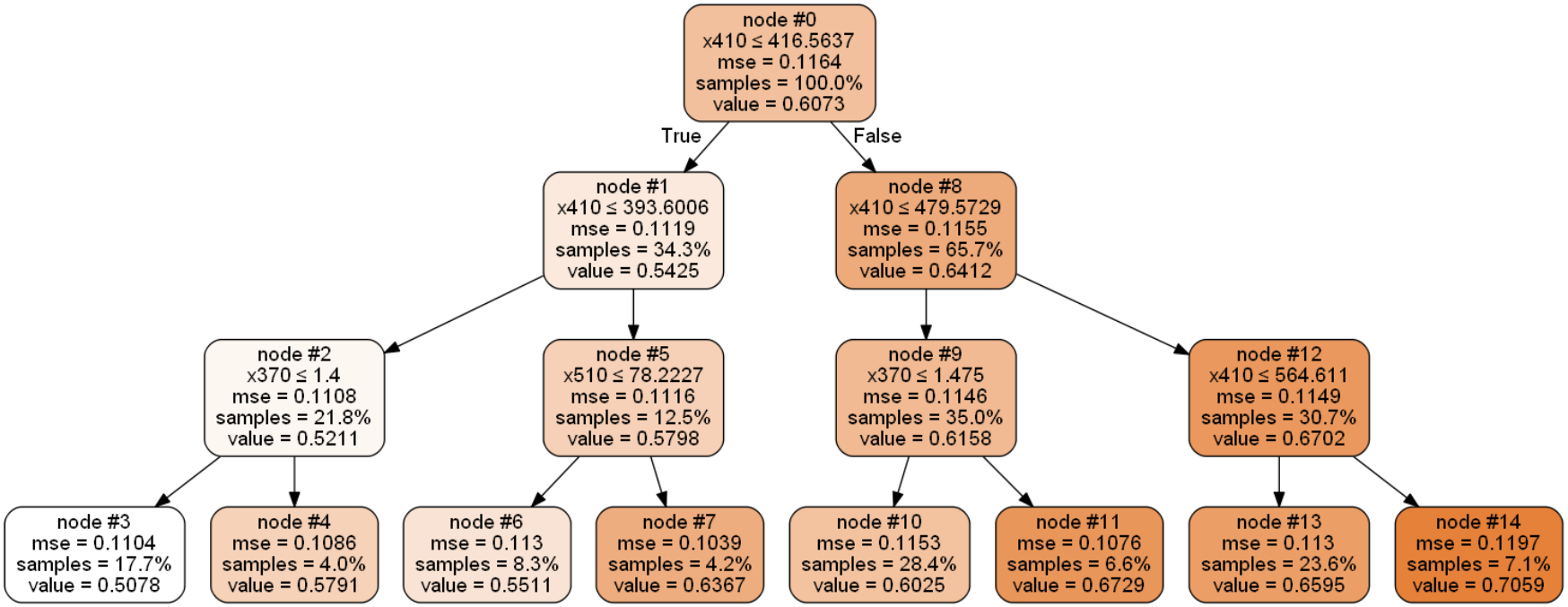
G.4.2 Decision tree diagram of transport mode Train



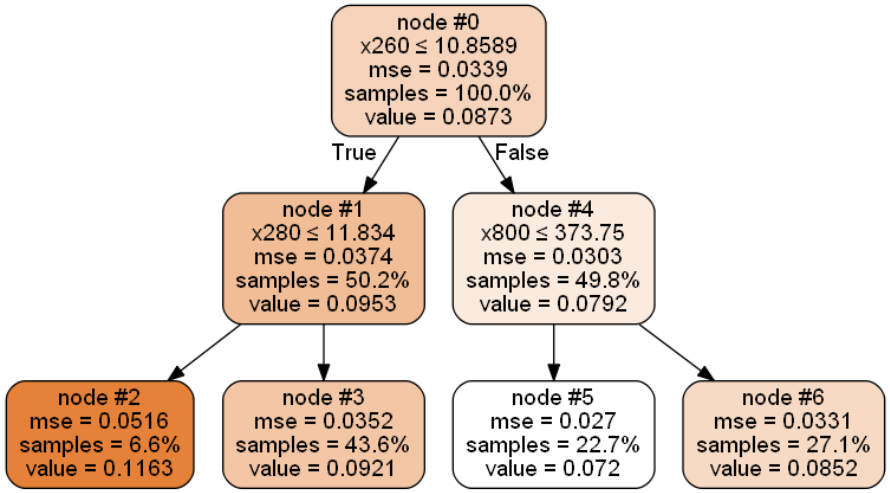
G.4.3 Decision tree diagram of transport mode MCT



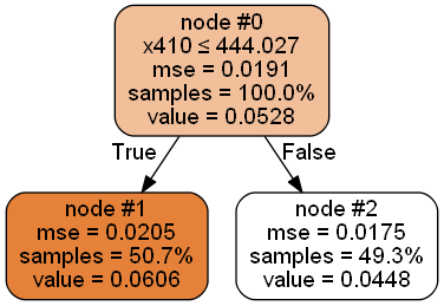
G.4.4 Decision tree diagram of transport mode Driver



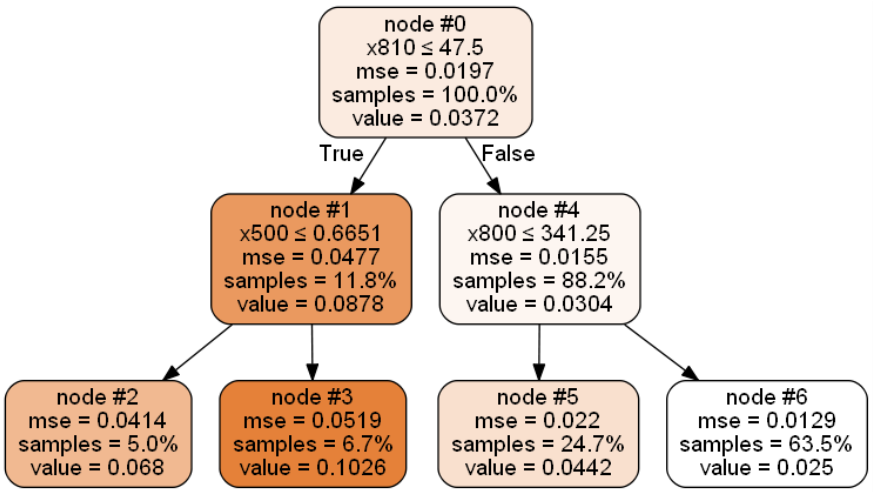
G.4.5 Decision tree diagram of transport mode CarPass



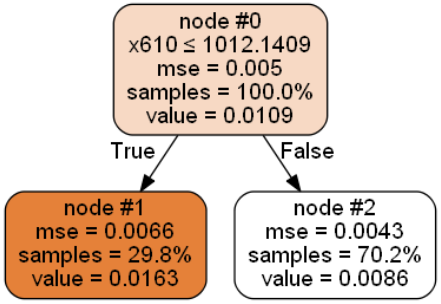
G.4.6 Decision tree diagram of transport mode PTplus



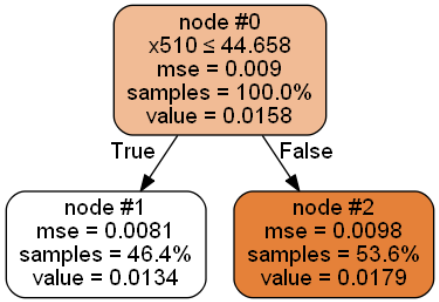
G.4.7 Decision tree diagram of transport mode Bike



G.4.8 Decision tree diagram of transport mode PTcom

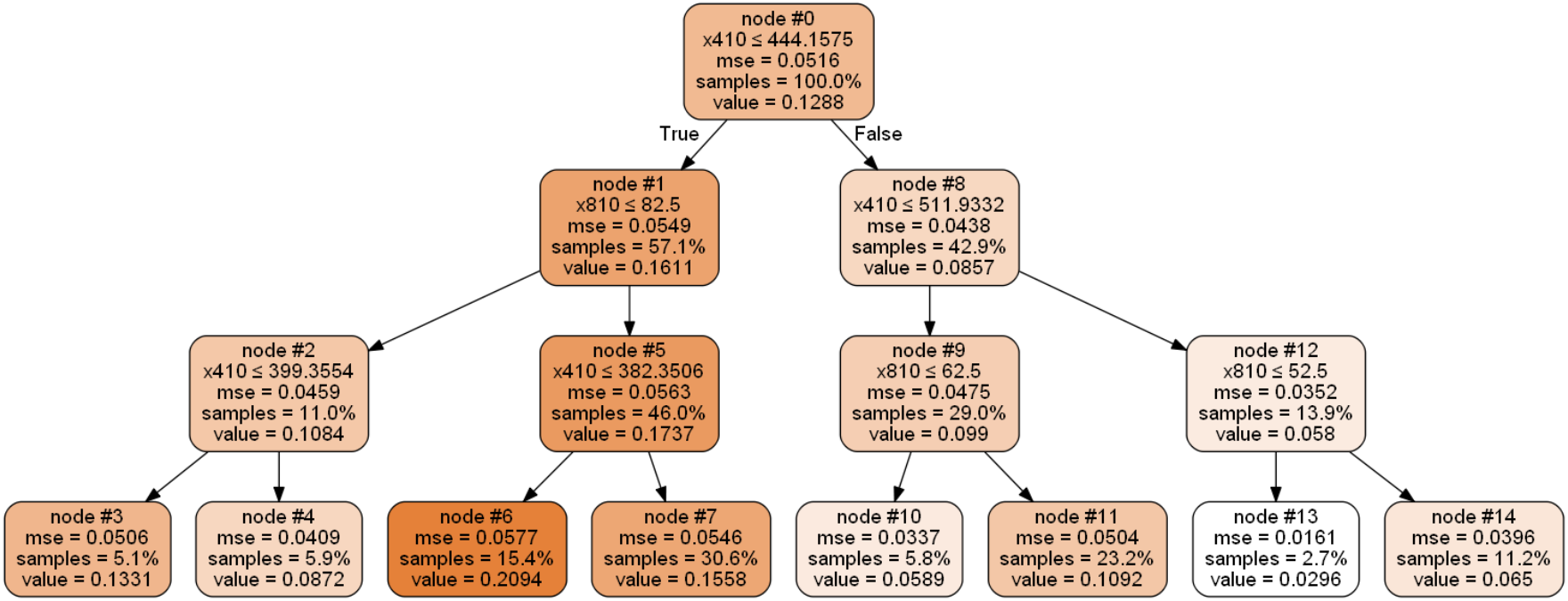


G.4.9 Decision tree diagram of transport mode Walk

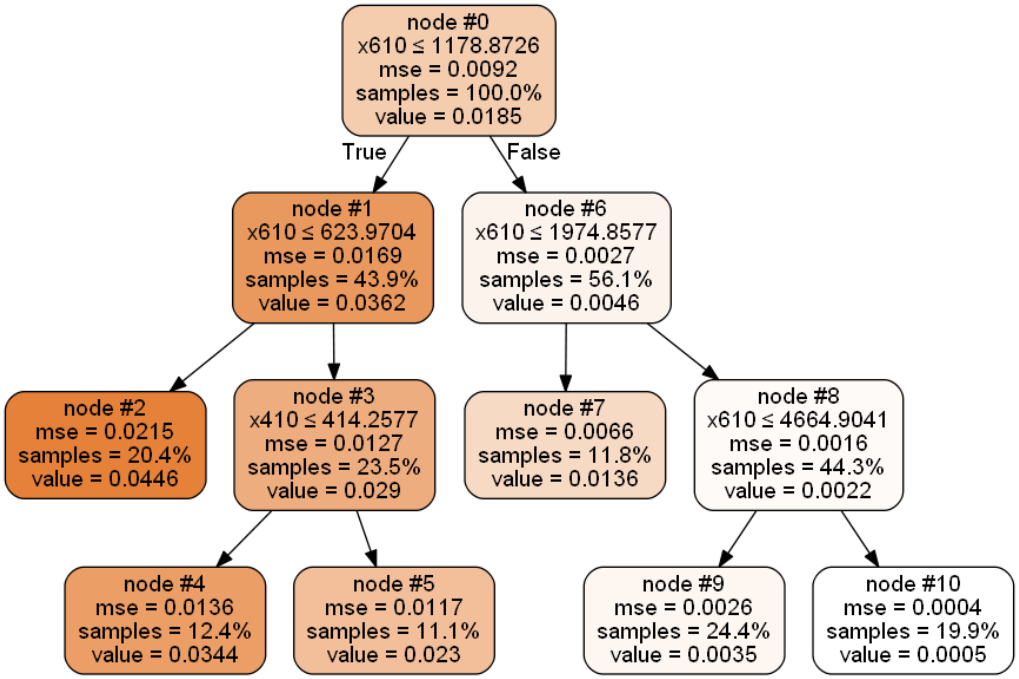


G.5 Type B models of work commute

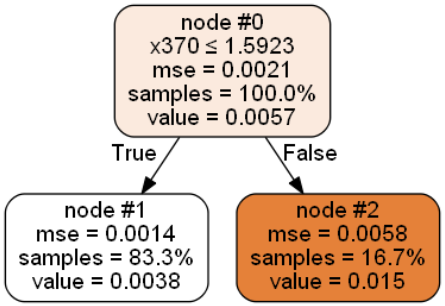
G.5.1 Decision tree diagram of transport mode Bus



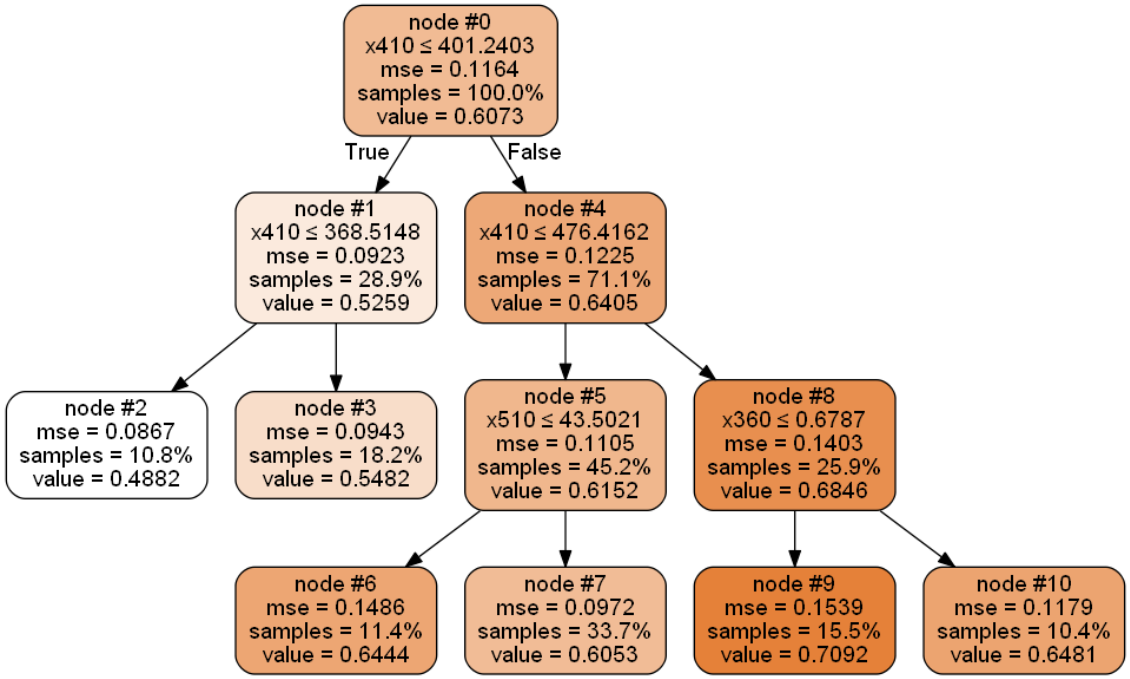
G.5.2 Decision tree diagram of transport mode Train



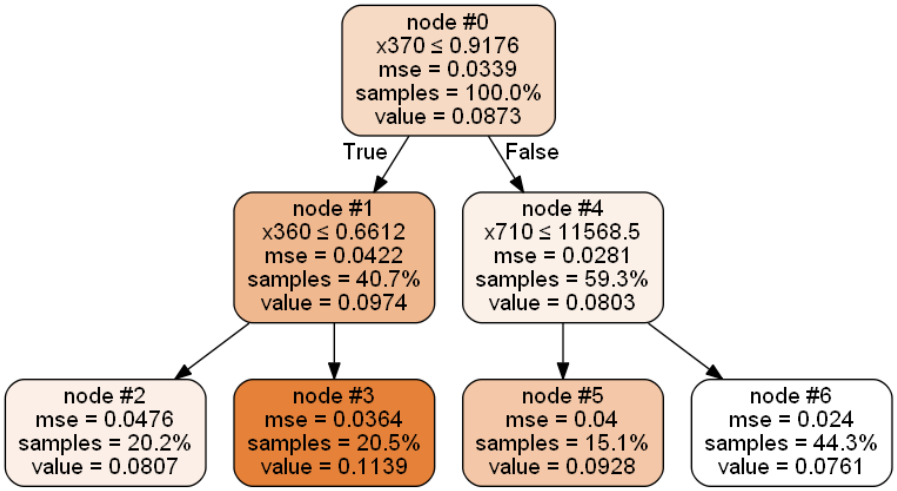
G.5.3 Decision tree diagram of transport mode MCT



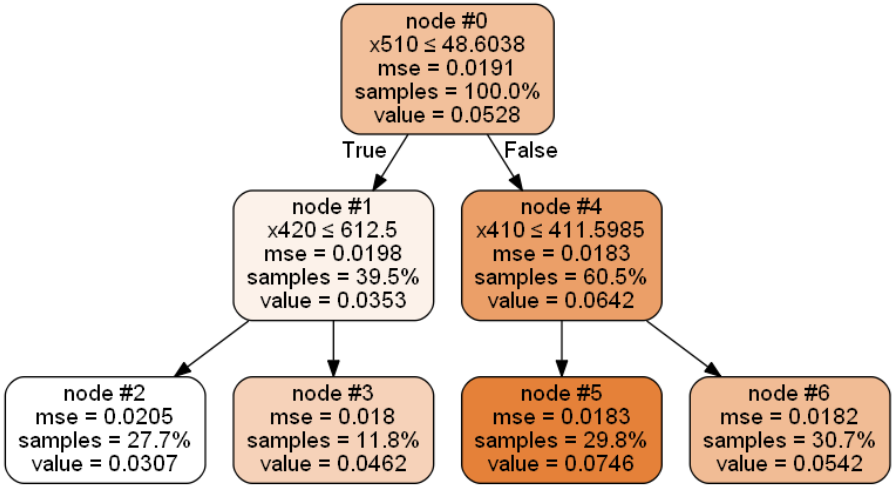
G.5.4 Decision tree diagram of transport mode Driver



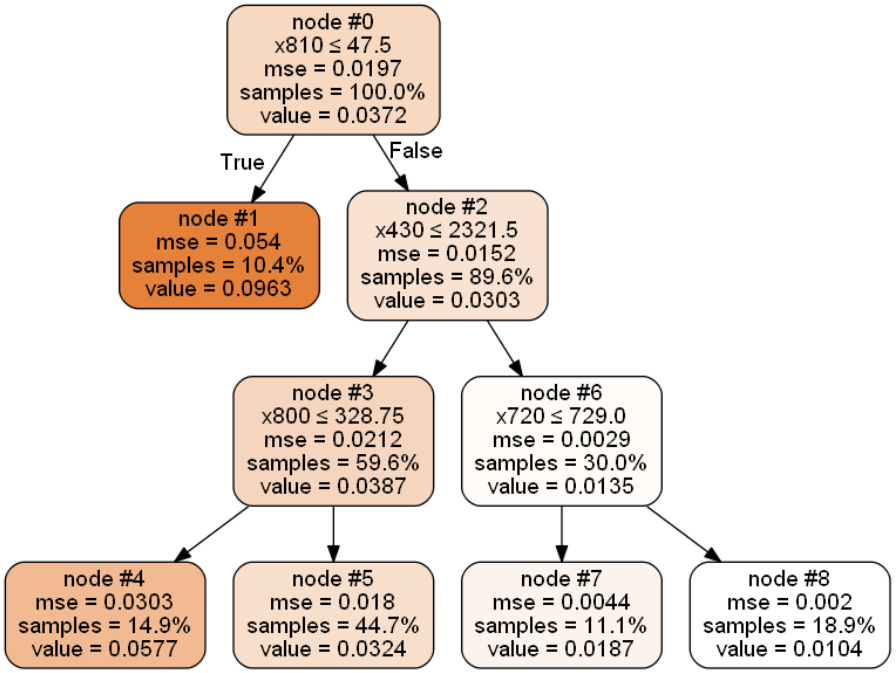
G.5.5 Decision tree diagram of transport mode CarPass



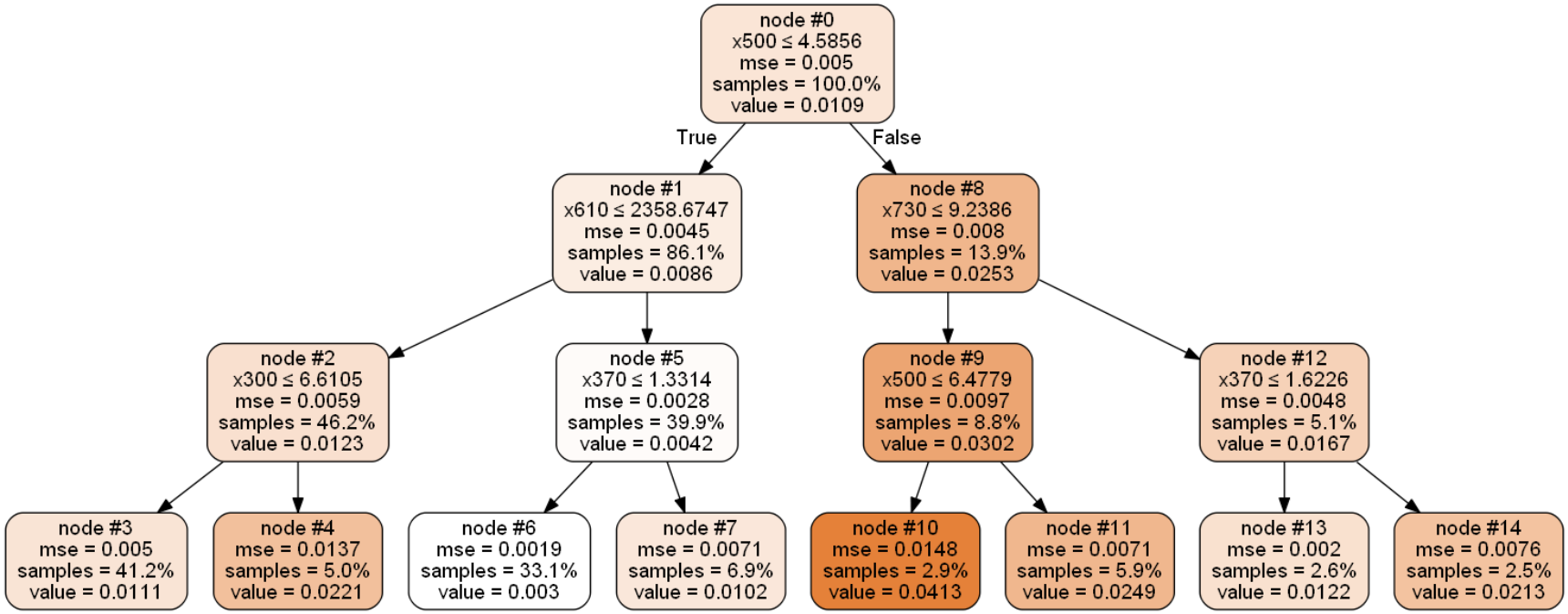
G.5.6 Decision tree diagram of transport mode PTplus



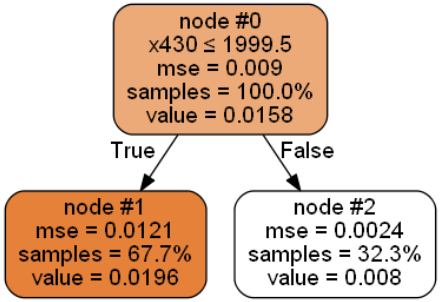
G.5.7 Decision tree diagram of transport mode Bike



G.5.8 Decision tree diagram of transport mode PTcom

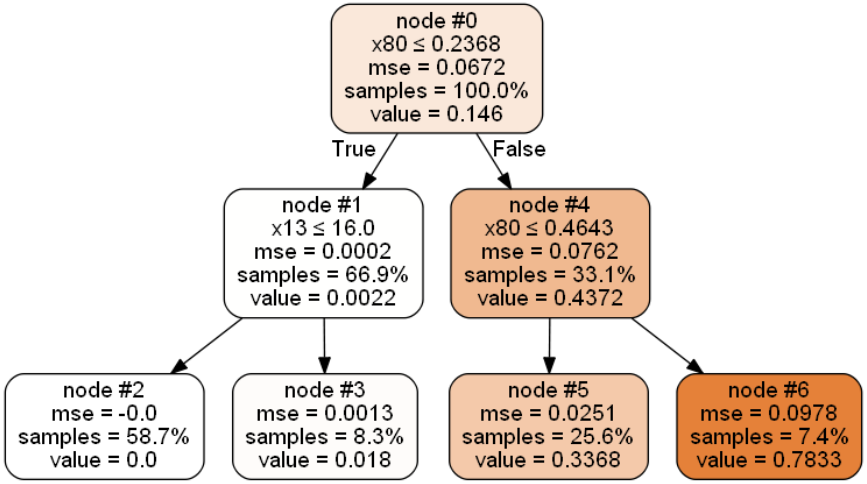


G.5.9 Decision tree diagram of transport mode Walk

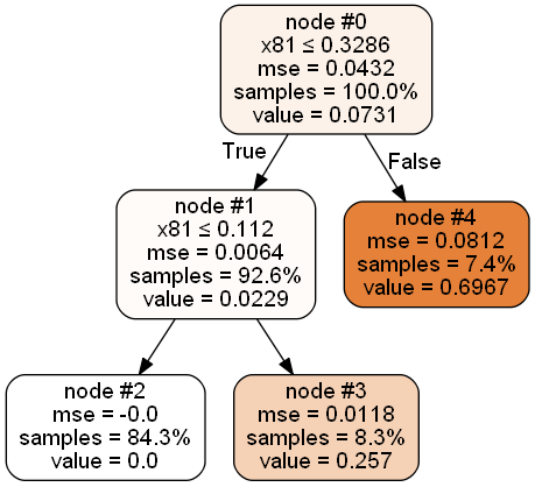


G.6 Type C models of work commute

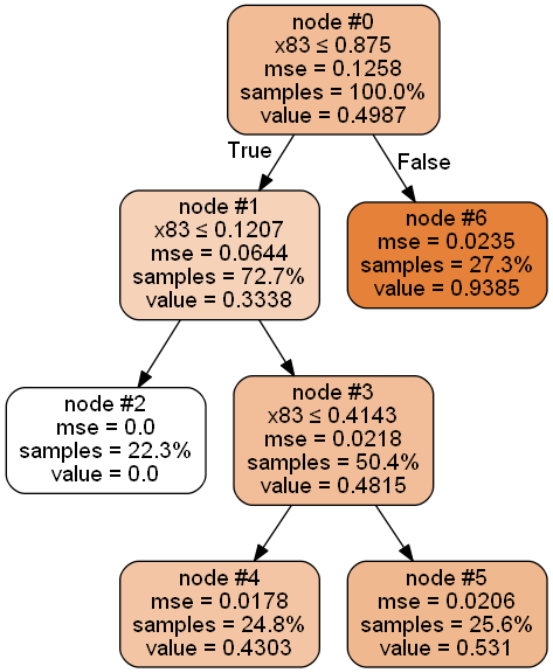
G.6.1 Decision tree diagram of transport mode Bus



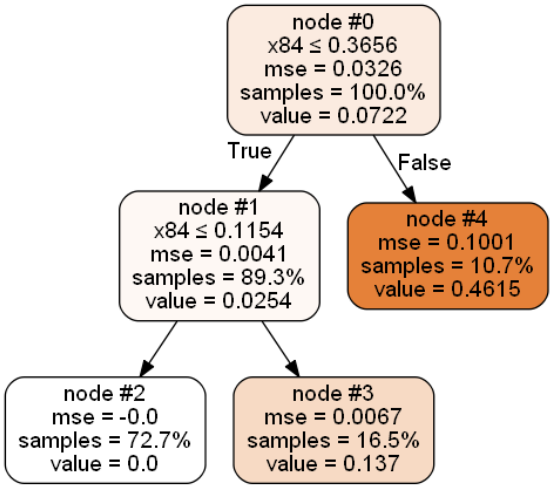
G.6.2 Decision tree diagram of transport mode Train



G.6.3 Decision tree diagram of transport mode Driver



G.6.4 Decision tree diagram of transport mode CarPass



G.6.5 Decision tree diagram of transport mode Bike