Aalto-yliopisto
Insinööritieteiden
korkeakoulu

Reetu Jormakka

Validation of mobile network data in producing Origin-Destination matrices

Thesis submitted for examination for the degree of Master of Science in Technology.

Espoo 22.11.2019
Supervisor: Claudio Roncoli
Instructor: Tapio Levä

Aalto University
School of Engineering

Aalto University, P.O. BOX 11000, 00076 AALTO
www.aalto.fi

Abstract of master's thesis

Author Reetu Jormakka

Title of the thesis Validation of mobile network data in producing Origin-Destination matrices

Master programme Spatial Planning and Transportation Engineering **Code** ENG26

Thesis supervisor Claudio Roncoli

Thesis advisor Tapio Levä

Date 22.11.2019 **Number of pages** 93 **Language** English

Abstract

The rapid development in telecommunication networks during last years has made it possible to study human travel behaviour effectively from mobile network data. The combination of passive and active signalling events gathered by mobile network operators allow analysing movements of people with full longitudinal and spatial coverage. Therefore, recent years have seen an increasing interest in utilizing mobile network data in transportation studies, as an alternative or a complementary data source for conventional transport data.

This study validates the capability of mobile network data to produce long-distance origin-destination matrices in Finland. Features that are being validated include trip counts, seasonal trip count changes and modal split. As reference data sources of the study, the National Travel Survey 2016, HELMET-transport demand model (Transport model by HSL) and LAM-data (automated traffic census) are used. Validation is done by analysing correlations between mobile network data and the reference data sources. By being able to demonstrate the validity and reliability of mobile network data usage in producing origin-destination matrices, cost-effectiveness and more accurate methods to gather information from long-distance transportation can be provided for the field in general.

The overall results of the study are in line with the few similar related studies that have been conducted. The thesis work suggests that mobile network data is capable of producing more reliable trip counts from sparsely populated areas than the National Travel Survey. In addition, it seems to be more capable of capturing the high summer peak in long-distance travelling in Finland. The results regarding modal split are promising, but more studies regarding the modal detection will be needed.

Keywords Mobile network data, Travel behaviour, Origin-Destination Matrix, ODM, The National Travel Survey

Aalto-yliopisto
Insinööritieteiden
korkeakoulu

Aalto-yliopisto, PL 11000, 00076 AALTO
www.aalto.fi
Diplomityön tiivistelmä

Tekijä Reetu Jormakka

Työn nimi Mobiiliverkkodatan käytön validointi lähtö-määränpää -matriisien luomisessa

Maisteriohjelma Maankäytön suunnittelu ja liikennetekniikka **Koodi** ENG26

Työn valvoja Claudio Roncoli

Työn ohjaaja Tapio Levä

Päivämäärä 22.11.2019 **Sivumäärä** 93 **Kieli** englanti

Tiivistelmä

Matkapuhelinverkkojen viime vuosien nopea kehitys on mahdollistanut yhä tarkemman matkapuhelinten solupaikannuksen. Teleoperaattoreiden keräämä passiivisten ja aktiivisten matkapuhelinverkon signaalihavaintojen yhdistelmä mahdollistaa ihmisten liikkumiskäyttäytymisen tutkimisen kattavasti sekä ajallisesti että alueellisesti. Viime aikoina matkapuhelinverkkodatan hyödyntäminen liikennetutkimuksissa on tästä syystä herättänyt kasvavaa kiinnostusta perinteisten tiedonkeruumenetelmien korvaajana ja täydentäjänä.

Tämä tutkimus validoi mobiiliverkkodatan käyttöä lähtö-määränpää -matriisien luomisessa Suomen pitkän matkan liikenteessä. Validoitavia ominaisuuksia ovat matkamäärät, matkamäärien vuodenajoittainen vaihtelu, sekä matkojen kulkumuotojakauma. Referenssiaineistona työssä käytetään Suomen henkilöliikennetutkimusta, HELMET-liikennemallia ja LAM-dataa. Validointi suoritetaan analysoimalla mobiiliverkkodatan ja referenssiaineistojen välisiä korrelaatioita. Osoittamalla mobiiliverkkodatan käytettävyys lähtö-määränpää -matriisien luomisessa, liikennesuunnittelun kustannustehokkuutta ja keinoja tarkemman tiedon keräämiseen pitkämatkaisesta liikkumisesta voidaan edistää.

Työn tulokset ovat linjassa aiemman tutkimuksen kanssa. Tulokset näyttävät mobiiliverkkodatan olevan kykenevä tuottamaan lähtö-määränpää -tietoa haja-asutusalueilta luotettavammin kuin Henkilöliikennetutkimus. Lisäksi, mobiiliverkkodata näyttää pystyvän observoimaan kesän lomakauden matkapiikin tarkemmin kuin Henkilöliikennetutkimus. Tulokset mobiiliverkkodatan kulkumuototunnistukseen ovat lupaavia, mutta lisää tutkimusta tarvitaan näiden havaintojen vahvistamiseen.

Avainsanat Matkapuhelinverkkodata, liikkumiskäyttäytyminen, Lähtö-määränpää -matriisi, Henkilöliikennetutkimus

Contents

Abbreviations

BSE	Best Server Estimate
CDR	Call Detail Record
CI-LAC	Cell ID – Location Area Code
CGI	Cell Global Identity
CRM	Customer Relationship Management
GSM	Global System for Mobile Communications
GPS	Global Positioning System
HELMET	Helsingin seudun liikenne-ennustejärjestelmän yksilömallit (Helsinki Region Traffic Demand Models)
HSL	Helsingin Seudun Liikenne (Public Transportation of Helsinki Region)
LAM	Liikenteen automaattinen mittausjärjestelmä (Traffic Flow Monitoring System)
NTS	The National Travel Survey
OD	Origin-Destination
ODM	Origin-Destination Matrix
OSM	Open Street Map
SMS	Short Message Service
SQL	Structured Query Language
VR	Valtion Rautatiet (Finnish State Railways)

1 Introduction

This chapter provides an introduction to the thesis work. First, overall picture of the topic is given in addition to the history and current state of the field. Then the thesis work project, regarding its objectives, framework and limitations are added in to the equation by showing how the thesis work contributes to the earlier studies and the future of the field. Finally, the structure of the thesis is provided to guide the reader through each chapter.

1.1 Background

On 1 July 1991, the former prime minister of Finland, Harri Holkeri, made a phone call from Helsinki to Tampere reaching Kaarina Suonio, the deputy mayor of city of Tampere. This was the world's first GSM phone call. Since then, it has been possible to locate a phone from the GSM network.

Whenever a mobile device is connected to a cellular mobile network, a signaling event is collected by the mobile network operator for billing purposes. An operator needs to know how long the phone call or an SMS-message was and where it was delivered in order to bill the customer accordingly. The fact that a mobile phone is always connected to the nearest base station, enables the mobile network operator to know where the mobile phone user is, at least in the terms of the base station location. However, this has not traditionally been very useful information.

When the development of the Global System of Mobile Communications (GSM) networks advanced and the increased amount of base stations were put up all over the world, also transportation researchers noticed the positioning possibilities of the GSM network. Roughly at the turn of the millennium, studies started to emerge in utilizing mobile network data in scientific research regarding movements of people (Streenbuggen et al., 2013). However, even though mobile network data was titled as a promising new data source in human mobility already two decades ago, its utilizations in production have not significantly emerged.

The major problems in commercialization of mobile network data in transport planning have been the inadequacy of the data accuracy and the relatively low profit potential of the data business for mobile network operators. All the way from 1991 to around 2010, GSM telecommunications relied mostly on phone calls and SMS-messages. Hence, signaling events were relatively rare and base stations were needed only sparsely. In addition, as remarkably large portion of the mobile network operators' (highly profitable) revenues were generated with subscriptions focusing on phone calls and SMS-messages, no incentives for alternative business models existed. However, today all of that is history.

Today, mobile phones are not just communication devices anymore. They are inseparable parts of our arms, providing us tools to work, ways to consume, abilities to function in the environment, contents for entertainment and networks for social relations. As everything happens in the internet, the telecom business has transformed into a business of providing internet connections. Suddenly, we have entered a new digital era, where "data is the new oil". The conventional telecom business is extremely highly competed and saturated, which affects the revenues negatively. This drives mobile network operators into new business models to ensure growth. As a sum of these variables, signaling events are created almost

constantly enabling the full potential of mobile network data in the field of transportation research, and what's even more important, the data is getting available for the first time.

With an increasing pace during last few years, studies from all over the world have emerged regarding the utilization of mobile network data in the field of transportation. Conventional data sources used in the transport planning, household travel surveys and traffic census, are expensive to produce and difficult to keep up to date. Their temporal and spatial representativeness cannot compete with the real time big data produced by mobile phones. Even as surveys and census have their roots deep in the transport planning traditions and processes, it is said that "they are not able to keep the pace of city growth and change, making relevant dynamic phenomena to be invisible for transportation and urban planners" (Graells-Garrido, 2018).

1.2 Purpose of the Study

This thesis work uses a set of mobile network data to study whether mobile network data can be used to produce reliable origin-destination matrices with trip counts, seasonal changes, modal split information and route choice information. By being able to demonstrate the usability of mobile network data in transportation planning, the whole field could benefit from increased effectiveness and lower costs. This regards especially public transport planning organizations which are funded with public money.

According to Bonnel et al., (2018) and Zagatti et al., (2018), the validity of mobile network data being used in producing origin-destination matrices has not been studied enough. By validating the capabilities of mobile network data in delivering accurate insights in human mobility, the overall development of transportation planning can be supported regarding a transition into new and alternative data sources. Like already seen in studies outside of Finland, the most obvious need for next-generation data lies in the field of transport demand modelling, especially regarding the replacement of national household travel surveys and field operated traffic census. The greatest motivation behind the usage of mobile network data in those operations is its remarkable productional cost-effectiveness compared to the conventional data sources. This fact forms the need of validation as follows: Is mobile network data able to produce equal (or better) results than conventional data sources, with remarkably smaller costs? And in addition: Is it possible that mobile network data could provide some insights even better than other data sources?

As it will be demonstrated in the latter parts of the thesis, mobile network data can be used in wide range of operations regarding transportation. For example, it can be used to generate origin-destination analyses, activity detections and traffic flow counts. This forms a limitation regarding the study, as not all of these applications can be validated in the thesis work project. For this practical reason, some of the most valuable assets of mobile network data were chosen to be validated, these being trip counts in long-distance travelling, route choices in long-distance highway traffic and transport mode detection in long-distance travelling. Based on these factors, the research questions of the thesis work are put as follows:

- Can mobile network data be utilized in long-distance transportation planning?
- Do the results of the mobile network data correlate with conventional reference data sources?
- Is mobile network data able to provide unique insights regarding seasonal changes, route choice or modal split, that cannot be acquired from other information sources?

Regarding earlier studies (e.g. Alexander et al., 2015; Bonnel et al., 2018; Graells-Garrido & Saez-Trumper, 2016), the hypothesis of the thesis work was that trip counts regarding long-distance transportation will turn out to be valid and reliable. Likewise, it was assumed that the result will be consistent with the origin-destination trip count results of the reference data sources, especially with the National Travel Survey 2016, which is the most comprehensive clarification of Finnish mobility and the main reference data source of the study.

However, based on earlier work (Huang et al., 2019), mobile network data's capability of providing unpresented insights regarding seasonal changes, route choices and modal split was assumed to be more difficult. Both route choices and the modal split would need more accurate location information generated from the signaling events than just trip counts between an origin and a destination, which increases the possibility of variance in the results. Especially distinguishing between modes of transportation has been considered traditionally as a weakness of mobile network data. At the moment, this requires individual consideration of each OD (origin-destination) pair and utilization of additional spatial data regarding transport infrastructure. Thirdly, and maybe most importantly, validation of these features is difficult due the lack of reference data. This however naturally demonstrates the demand for alternative data sources regarding modal split, route choices and seasonal changes, as this kind of information is not available with current data gathering methods.

The contribution that this thesis work tries to bring for the research community of the field lies in bringing varying set of reference data sources to the validation process, and incorporating different mobile network data analyses, such as trip counts, temporal trip count changes, modal split and route choice, into one combination of analyses. Additional elements of the thesis work that stand out from other related studies are the type of the data that is being used and the commercial starting point behind the data production. Unlike with most of the related studies of the field that have utilized a relatively small set of CDR (Call Detail Records) data for their analyzes, the combination of passive and active signaling data enables conducting large longitudinal origin-destination matrices from huge and accurate source data. For demonstrating the scale, the great regional ODM analysis of the thesis work is generated from movements of 1,8 million Finnish people during a year, which makes a total of 1,77 billion trips, which are then again created from roughly 100 billion rows of signaling event data. This is an amount of source data event points that has not probably been seen in the related studies and provides this way possibly new insights into the capabilities of mobile network data overall.

1.3 Framework of the study

So far, Telia is the first and the only mobile network operator which has commercialized mobile network data in Finland and brought it available for other companies and public organizations. Hence, this thesis work was done in a collaboration with Telia Company, which operated as the supplier of the thesis work by providing the mobile network data, an

IT-infrastructure and algorithms for the analysis. In addition, Strafica Oy (currently Ramboll) and WSP Oy acted as field specific consultants and a support regarding transportation. This made it natural to build the case study to incorporate geographically only Finland to narrow down international variables.

The thesis work study will follow the guidelines of earlier researches regarding mobile network data validation and consider additional elements from studies involving modeling of modal split and trip assignment. Everything about the study is tied into trips between known origins and destinations, and no other forms of data are considered (for exception, some traffic flow counts are used as reference data). The core of the study is this way detecting trip counts between given origin-destination pairs and then deepening the understanding of those trip counts by classifying the trips in multiple ways.

Long-distance trips were chosen to be the core of the analysis, as they are difficult to capture with conventional traffic data. Hence, long-distance trips are usually also the portion of trips which are most poorly known by transport administrations and most difficult to be captured with household travel surveys (Bekhor et al., 2013). When the effects of some ongoing megatrends are considered, such as negative externalities of air travelling and positive externalities of rail travelling, it becomes significantly important to know how, where, when and how much people perform long-distance trips in order to support the development of sustainable transportation.

The National Finnish Travel Survey 2016 (Finnish Transport Agency, 2018a) is the main reference data source of the study. This is the most comprehensive data source regarding Finnish mobility, as no other Finnish data source contains as large number of origin-destination trips from all over Finland. In addition, other data sources such as LAM detectors (Finnish Trasnrport Agency, 2019a) and the HELMET-transport demand model (HSL, 2014) are used as reference.

The project started by creating OD-matrices both from Telia's mobile network data and from reference data sources. First, only trip counts were considered. Trip counts from mobile network data OD-matrices were compared to OD-matrices from other data sources, and correlation between these matrices was examined. After a correlation was acquired (or in a single case was not acquired), more moving parts were added into the equation by classifying the trip counts between OD-pairs by transport modes, route choices and travel durations. Then, the correlations were examined again. The aim of the validation was to demonstrate that Telia's mobile network data matches with other reliable data sources. In case of a match, a proof was found that showed the potential of mobile network data being used to acquire data from traffic extremely cost-effectively. However, the reliability of the reference data sources was also put under question in some cases.

The scope of the study does not include development of Telia's mobile network data processing algorithms or testing different solutions to create better processes and this way more accurate results. Instead, it focuses purely on using the already existing solutions and tries to validate them by utilizing knowledge regarding spatial planning and transportation engineering. However, all observations are naturally reported publicly as thesis work results to enable possible improvements regarding the usage of mobile network data in the telecom and transportation sector in the future.

1.4 Limitations of the Research Scope

As the validation and analysis was done only with Telia's mobile network data, no clear assumptions can be made that would promote mobile network data usage regardless of the algorithms or the telecom operator. This means that as the algorithms and other moving parts are inseparable part of the Telia's data production pipeline, not only the mobile network data results are validated but also the collaboration of algorithms, data warehouses, telecommunication networks, customer databases and other functions. So basically, what was validated was Telia's capability to provide insights in human mobility, not mobile network data generally.

Some deeper level understanding could have been acquired by requesting the raw signaling event dataset as a single data delivery and then by developing all of the data processing algorithms in the name of this particular research. However, that would not have been possible regarding the thesis work resources. In addition, having all the algorithms already functioning in production was the very reason why it was possible to access the vast amount of mobile network data flexibly with different kind of queries, aggregations and timeframes during only few months.

1.5 Structure of the Thesis

The thesis starts with a theory chapter, which introduces the reader to the fundamentals of mobile network data positioning. This chapter describes how does a mobile network infrastructure function, what different methods are there for a GSM network to locate a mobile phone, what kind of different signaling events are created and how trips are generated, aggregated and extrapolated from the signaling events. In addition, the chapter gives a short introduction into theory of household travel surveys, origin-destination matrices and related work of the field.

Methods and materials of this thesis work are not presented in the same chapter, as separating these entities into different chapters makes it clearer to demonstrate what was done in the actual thesis work project and what has been made by other organizations already before the thesis work project. The major structural idea is that the things that were there already before the thesis work project, like the results of the National Travel Survey 2016 and Telia's data processing algorithms, are presented in the Data & Data Processing -chapter. This is natural, as these things work as starting points of the thesis work; in a way as materials that are being used to generate the study results.

The Data & Data Processing -chapter starts by describing how does Telia produce its OD-matrices from the mobile network data. Basically, the same aspects are went through that were already seen in the theory chapter, but this time only from the viewpoint of Telia. Then, the National Travel Survey 2016 is introduced in full detail and the extrapolation of the survey participants is discussed.

After the Data & Data Processing -chapter, the Methods -chapter describes how these two data sources were refined to origin-destination matrices and then analyzed. This chapter demonstrates the parameters given for the OD-matrices and the tools and methodologies that were used to compare these results. Basically, what was done to the data sources presented in the Data & Data Processing -chapter to create the results which are presented in the Results

-chapter. Concepts of consistency ratios and Pearson correlation are introduced, as these are the methods that were used to present the reliability and validity of the mobile network data in the Results -chapter. In the beginning of the Methods -chapter, a project flow chart is presented to demonstrate the project structure more clearly (Figure 8., section 4.1).

The Results -chapter presents the results of the analyses. The correlation and consistency between Telia's mobile network data and the reference data sources is presented with correlation ratios and consistency matrices. In addition, mobile network data's ability to profile seasonal changes, route choice detection and modal split information is presented. Discussion and Conclusion chapters sum up the findings of the thesis work.

2 Theory

2.1 History and applications of mobile network data

Recent years have provided fast-growing amount of studies regarding mobile phone data usage in the field of human mobility. As cities are growing fast in addition to developing mobile technologies and increasing phone usage, mobile phone data becomes more relevant data source for movements of people year by year. The usage of mobile phone data has been studied already for twenty years with accelerating rate, and during the last few years the pace has even increased. Hence, related work, different applications and use cases are countable in hundreds and for this reason not all of them are presented here. However, the most common and notifiable use cases regarding this thesis work are tried to be presented in this chapter.

Starting with studies that do not specifically go into the field of transportation engineering, other topics, such as social networks, activity spaces, ethnic distribution, and healthcare can be mentioned for example. Mobile network data has been used to study human mobility related to people's social ties' locations, showing how people locate compared to their social contacts (Phithakkitnukoon et al. 2012). Järv et al., 2015 studied spatial distribution of people with different ethnical backgrounds, and network of public healthcare facilities compared to people's movements was assessed by Wesolowski et al., 2015. These studies demonstrate that mobile network data can effectively be used also in the fields of sociology and communications.

At the population level and with larger geographic areas the range of use cases goes even wider. Due the aggregated and large sample sized nature of the mobile network data, applications related to temporal and spatial distributions and dynamics are studied widely. Similarly to ethnical distribution of people, also the distribution of local and national population in cities has been studied with mobile phone data (Deville et al., 2014). Due the longitudinal elements of mobile network datasets, seasonal migration patterns (Silm & Ahas, 2010) and tourism (Ahas et al., 2008; Nurmi, 2018) are use cases where mobile data has also been applied widely. These studies demonstrate how mobile network data can give remarkable insights about human masses moving between areas for example regarding different seasons and holidays. One obvious use case is spatial planning and land use (Louail et al., 2014), in addition to socio-economic levels and their distribution within a city (Blumenstock et al., 2015). Wesolowski et al., 2012 introduced a use case also for the field of health geography, regarding disease risks.

In the field of transportation engineering, mobile phone data has been recognized as a valuable raw data option for many kinds of applications already for a couple of decades, and one of the first large scale mobile phone data studies in the field of transportation was conducted by Gonzales et al., 2008 with sample size of 100 000 individuals. In the transportation field the most revolutionary application has probably been the ability to infer human travel patterns from mobile network data (e.g. Asakura & Hato, 2004; Phithakkiitnukoon et al., 2010; Phithakkitnukoon et al., 2011; Reades et al., 2009; Widhalm et al., 2015). This means the ability to derive mobile phone user's trips and activities throughout different timeframes and this way understand where the user lives, works, shops, spends time and how and when does the user travel between these locations. Also, a wide range of studies have extracted wider mobility patterns based on individual users and this

way applied these travel patterns into for example road usage, transport networks, origins and destinations and travel purposes (e.g. Asgari et al., 2013; Candia et al., 2008; Gonzales et al., 2008; Simini et al., 2012; Sevtsuk & Ratti 2010; Song et al., 2010; Wang et al., 2012).

Also more specific methods of utilizing mobile network data have been developed into the field of transportation engineering. For example, route choice modelling with mobile phone data has given promising results compared to conventional route assignment methods and models (e.g. Becker et al., 2011; Schlaich et al., 2010). Actually, the term "modelling" regarding route choice extraction from mobile phone data might be little misleading, as mobile phone data records directly show the base stations used for connections and this way also the route used. Transport models can be calibrated directly with mobile phone data instead of travel surveys (e.g. Bolla et al., 2000) and traffic flows have been estimated effectively from mobile phone datasets (e.g. Cheng et al., 2006; Demissie et al., 2013). In addition, recent years have shown promising steps towards more accurate modal split estimation, even though this element has traditionally been considered as the weakness of mobile network data (e.g. Bachir et al., 2019; Calabrese et al., 2010; Graells-Garrido et al., 2018; Qu et al., 2015).

Similarly to this thesis work, the validity of mobile network data in producing OD-matrices has been studied already in some extent (Alexander et al., 2015; Bekhor et al., 2013; Bonnel et al., 2018; Chen et al., 2014; Graells-Garrido & Saez-Trumper, 2016; Iqbal et al., 2014; Kujala et al., 2016; Zagatti et al., 2018). Studies by Graells-Garrido & Saez-Trumper, (2016) and Bonnel et al., (2018) used correlation and regression to examine the validity of mobile network data against household travel surveys.

2.2 Positioning capabilities of cellular mobile network

In addition to enabling calls, SMS-messages and data, mobile phones can be used for tracking locations. Mobile phones can be tracked for example through GPS (Global Positioning System), GSM (Global System for Mobile Communication) and RF-ID (Radio Frequency Identification), which all of them provide a wide scope of potential advantages compared to travel surveys and census in human mobility research (Asakura & Hato, 2004). However, only GSM is considered in this thesis work. The aim of this section is to present the basics of cellular mobile network (GSM) positioning capabilities and technologies whereas the advantages and disadvantages of different transport data sources are discussed in the section 2.12.

Functionality of locating a mobile phone through cellular mobile network has existed as long as mobile phones have been used, but only the technological development of last decades has made it more regular to utilize mobile networks for location-based services. There are several methods how to locate a mobile phone from a cellular network which all base more or less on few fundamentals. We will start by going through the very basics of GSM network and the fundamentals of its positioning capabilities. Then we'll distinguish between event-driven and network-driven mobile phone location notifications and open up Cell Global Identity (CGI) property of GSM network, which is the most basic way to obtain location information from the GSM network. After that we'll see shortly how one can obtain even more accurate location information than with CGI, and how then the users and trips are assigned spatially based on the location information received. The final sections of the

chapter present extrapolation, privacy and conventional data sources compared to mobile network data in the field of transportation.

2.3 Base Stations and Cells

When mobile phone (MS) is turned on it is connected to a nearest base station (BTS) which provides the network signal and receives and delivers data to the phone (Janecek et al., 2015). Typically, mobile network base stations carry multiple antennas directed to several directions, and coverage areas of these antennas are called "cells", meaning that one base station usually provides multiple mobile network cells (Ahas et al., 2008). Cells are the smallest and the most accurate spatial entities in the cellular network (Janecek et al., 2015), but their spatial coverage area is not fixed (Järv et al., 2017). Size of the cells are varied according for example the density of base stations (for some exceptions, see Calabrese et al., 2013), meaning that in cities the cells are typically smaller than in rural areas. Multiple base stations are controlled through fewer base station controllers (BSC) which are then linked with Mobile Services Switching Centers (MSC) that provide communication channel between two mobile phones registered in two different cells (Shad et al., 2012).

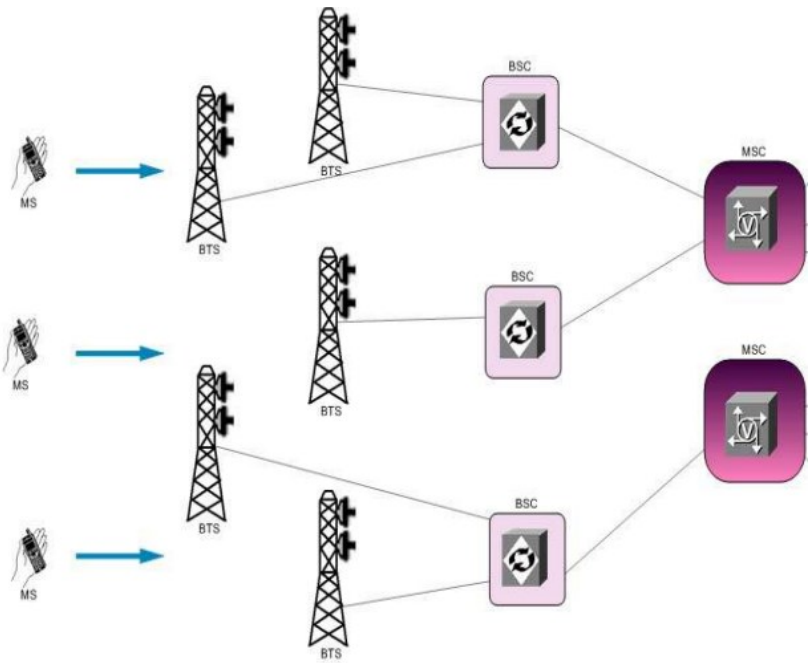


Figure 1. Architecture of functional GSM network (Shad et al., 2012)

A mobile network operator company has naturally knowledge of its base station coordinates, and it can determine the location of individual cells with omni-round radio coverage and radio coverage sectors determined with bearing degrees (Ahas et al., 2008). This way the operator has a solid knowledge of the regions covered by each cell and the cell boundaries. As mentioned, these coverage areas can vary in coverage significantly; in urban areas one might reach a cell coverage area as small as few hundred square meters, when sometimes in the peripheric areas they might be even 35 km long (Shad et al., 2012). (See section 2.5 for more detailed information about positioning accuracy.) When a mobile phone signal connects to the base station, the mobile phone notifies its position in the terms of the cell where it is currently located (Calabrese et al., 2015). This means, that without any kind of

higher-level software or specific system implementations, the operator is always aware of the Cell ID where the mobile phone is located when the phone is connected to the network (Shad et al., 2012).

2.4 Event-Driven and Network-Driven Mobile Network Data

The notification of a mobile phone connecting a base station can be generated by *events* (calls, SMS, data transfer) or by *updates of the network*. Basically, the difference between event-driven and network-driven notification is that the event-driven notification is made by user's initiative and network-driven can be made also by the network (Calabrese et al., 2015). Whenever a connection is created, the mobile network system typically saves at least the User ID (who owns the phone), Cell ID (the cell where the phone is located) and a time stamp (time) of the connection (Janecek et al., 2015), due the relevance of this information regarding telecom business. The most widely used event-driven mobile phone network data type in transportation engineering studies is called Call Detail Records (CDR), here mentioning a few related studies (e.g. Dong et al., 2015; Alexander et al., 2015; Gonzales et al., 2008). Call Detail Records are collected in most of the world's mobile networks (Ahas et al., 2008), as they provide the base for user billing. Even though each operator determines what data to include in CDR data scheme, CDR data can be almost exclusively used in urban sensing as time stamps and location IDs are the fundamental parts of the records. CDR records can also be used to study human social networks as they provide information about the respondent of the call/message. Another event-driven mobile network data type is called Internet Protocol Detail Records (IPDR) which contains details about user's internet usage (Huang et al., 2019). IPDR records contain additionally information about the bytes transferred and the websites visited. As event-driven records are created only through network events launched by the mobile phone user, they do not provide capability of constant locating.

Table 1. Example of a CDR log: a) originating and terminating user id, originating and terminating cell id, timestamp and call duration. (b) Cell location information (Calabrese et al., 2015)

originating_id	originating_cell_id	terminating_id	terminating_cell_id	timestamp	duration
24393943	10121	17007171	10121	24031517	29
24393943	5621	17007171	2721	25141136	38
24393943	17221	17007171	2521	25534630	188
24393943	31041	17007171	5111	32440483	111
24393943	10121	17007171	9411	33152308	145
24393943	6321	17007171	20921	33431903	132
24393943	7041	17007171	10021	33435718	17
24393943	7021	17007171	14321	34160370	53

(a)

cell id	lat	lon
10121	44.658885	10.925102
17221	44.701606	10.628872

(b)

As event-driven mobile phone network provides only "semi-constant" locating capability of mobile phones, the network-driven mobile phone locating systems are capable of locating the phones as many as 500 – 1000 times a day (Jiaqi, 2018). Network-driven mobile phone connections can be classified into *passive* events and *active* events. Network-driven active and passive events provide much larger volume of spatially coarse-grained mobility data and

finer-grained spatial accuracy than event-driven devices (Janecek et a., 2015) The passive mobile phone datasets are created regardless the user uses the phone or not (Ahas et al., 2008). There are three types of passive mobile phone network events:

1) Periodic Update, which provides the current Cell ID information of a phone on a periodic interval.
2) Handover, which is generated when a mobile phone moves from a cell to another, or from different telecommunication infrastructure to another (2G, 3G, 4G, 5G).
3) Mobility location update, which is generated when a mobile phone moves from a Location area to another. (Location area is a group of base stations which is managed through a single base station controller (Calabrese et al., 2015; Shad et al., 2012))

Active mobile positioning data is then again generated when a mobile phone is used by the user, but unlike with CDR, network-driven active records do not store information about the respondent of the connection (Huang et al., 2019). In general, we can see that network-driven datasets are year by year becoming more temporally frequent and spatially accurate, as the usage of mobile internet increases and GSM infrastructures develop (Huang et al., 2019). It is still also important to notice that gathering of the active and passive signaling events is not a by-product of the conventional telecom business but requires system implementations to the mobile network that are specifically designed for the purpose (Ahas et al., 2008). This means that unlike with event-driven data, the mobile phone operator does not acquire active and passive network-driven mobile phone data without investing into specific technologies.

2.5 Cell Global Identity (CGI) and other location detection methods

The basic idea in locating mobile phones from GSM network is that any cellular network is able to identify the current Cell ID of the mobile phone by recognizing the base station and antenna which the phone is using for the connection. This is called Cell Global Identity (CGI). CGI is the easiest way of gathering location information of mobile phones from the GSM network, as it is always readily available, and the Cell ID number is gathered automatically (Shad et al., 2012). As CGI is only aware of the cell where the phone is, the locating accuracy is directly proportional to the size of the cell, and without further technologies the operator is unable to specify the location of the phone within each cell.

The most relevant parts of information headers within CGI signaling data are Cell Identifiers (CI) and Location Are Codes (LAC) (Shad et al., 2012). With CI-LAC combinations operators can uniquely determine the centroid of the cell or coordinates of the base station within a mobile phone cell. The boundaries of the cell are then again determined by omni-round coverage and radio coverage sectors with bearing degrees (Ahas et al., 2008). The cell network estimation made by the operator is called Best Server Estimate (BSE), and it gives detailed information about the coverage areas of each mobile network cell and positioning of each antenna. Hence, BSE clearly divides the whole country into coverage areas (2G, 3G, 4G, etc.), that can be used for estimating which cell the mobile phone uses for network connection regardless of the location of the phone in the country. This information is usually used for the network coverage planning and design.

However, there are several additional methods to improve the accuracy of the cellular locating by determining the mobile phone locations within cells. For example, angle of

arrival (AOA), time difference of arrival (TDOA), finger point, ray trace, signal strength (SS), and GPS combined can be mentioned (Asakura & Hato, 2004). Some of these methods require the usage of multiple base stations (triangulation) and some of them only one, but in common for all of them is that in order to function they require upgraded mobile handsets or extra equipment installations to the network system (Shad et al., 2012).

Without getting too much into detail, couple of these methods can be introduced here. Time difference of arrival (TDOA) and signal strength (SS) base both on triangulated measurements between multiple base stations. Much like Global Positioning System with satellites (GPS), time difference of arrival measures the time spent by the signals departing from a mobile phone and arriving to the base station receivers (Xu et al., 2013). By knowing the signal travel times spent at least to three base stations, the location of the phone can be derived within a cell. Same methodology can be applied to received signal strengths of the phones. As the signal strengths at the start of signal transmission are well known and power drop during the signal transmission is well defined, the phone location can be calculated by measuring the signal strengths at least in three base stations (Fayaz et al., 2014). There are also methods such as propagation models and irradiation diagrams which can increase the accuracy of results by minimizing the mean square errors between measured and estimated values received at the base stations. With this kind of techniques, one can reach an accuracy level of around 150 meters regardless of the cell size of the current connection (Calabrese et al., 2015).

2.6 User Assignment

Giving the fact that Cell Global Identity is the most widely used location detection method in cellular network (Calabrese et al., 2015), but it cannot provide the accurate location of the mobile phone user, we will next go through different methods on how to assign the users to the real geographical areas. As with CGI we are only aware of the Cell ID where the phone is located, we cannot specify whether the user is in an office building located at a cell or at the nearby forest located at the same cell. There are several methods used in population research and mobile network research to interpolate and predict where people are located within a predefined census tract or this case a mobile network cell (Järv et al., 2017). The aim of these methods is to transform values from a certain spatial unit into another spatial unit. In this case from source units (mobile network cells) into target units (grid network, postal areas, municipalities, etc.) which can be more easily used in population research than mobile network cells and then be combined and validated with other spatial data sources with same boundaries (Järv et al., 2017). State of the art works in the area suggest few different main methods for the problem, basing on *area proportional assignment, balanced assignment* and *dasymethric assignment* (Girardin et al., 2009; Gonzales et al., 2008; Järv et al., 2017). In addition, one can distinguish a class of methods regarding user assignment called *probabilistic user assignment*, which tries to evaluate whether the Cell ID received from the log file can be trusted (Traag et al., 2011; Zang et al., 2010).

At the most basic level, area proportional user assignment means that users are assigned to spatial units based on the Cell ID of their network event log file. For example, if we want to examine population movement of our mobile network data with a 500 m x 500 m grid (target zones), the grid is put on top of the cellular coverage area (source zones), and the observations are then distributed into the smaller spatial entities, into the grids (Trasarti et al., 2015). Usage of different sized grids in the mobile network mobility research are

discussed in the study of Williams, (2015). The distribution can be done for example with a simple areal weighting method, see Fig. 2.

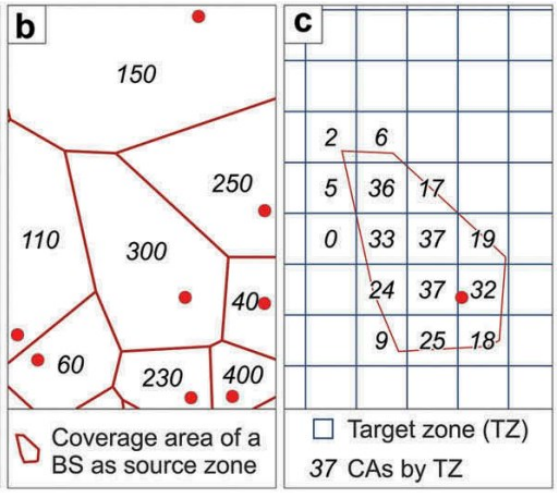


Figure 2. b) Base station coverage zones (or cells) and the number of mobile phones at a given time in each of the cells. c) Grid is put on top of the cell and the observations are distributed spatially into the grids with a simple areal weighting method (Järv et al., 2017)

In case of only point based source data, the cellular coverage areas can be derived using Voronoi tessellation (Gonzales et al., 2008), see Fig. 3. This means dividing a space in a way that each point is circled with a polygon that holds the areas in it which are closest to that specific point. Usage of Voronoi tessellation might be the case for example if the operator is not able to provide its cell boundaries or if the researcher prefers point based raw base station observation data. Problematic element of Voronoi polygons is that many objects of the urban infrastructure can interfere with the real coverage areas (Girardin et al., 2009). For example, buildings might block signals effectively even though they are located in the close proximity of base stations, but Voronoi polygons are placed without taking this into account. In most cases telephone operators are aware of their cell boundaries (overlapping for 2G, 3G, 4G and 5G), as the network coverage planning and design bases on that information (See section 2.5).

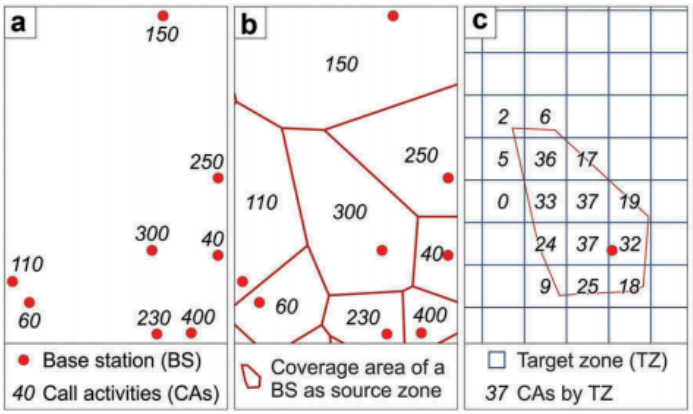


Figure 3. a) Observations are conducted only for base station level and then the cellular boundaries (b) (or base station coverage areas) are derived with Voronoi tessellation method (Järv et al., 2017).

There are several widely recognized challenges related to the area proportional user assignment. Firstly, the distribution of base stations is not spatially uniform as base stations are much more densely located in urban areas compared to rural (Williams et al., 2015). This creates a problem with peripheric areas where cell sizes may be several square kilometers and users this way assigned inaccurately. In addition, base station locations vary over time as urban infrastructures develop (Williams et a., 2015). This makes it challenging to compare datasets temporarily between years. The most problematic part of area proportional user assignment is probably still that it assumes space as a plane surface and people evenly distributed into that surface, without considering the effects of for example lakes, seas and forests, where people usually don't tend to spend time in reality (Järv et al., 2017).

The problem of different land-use types can be corrected with balanced user assignment. Balanced user assignment means that unlike with area proportional user assignment, the different land-use types (lakes, parks, forests, etc.) are considered in the assignment process within mobile network cells. Simplified, this means for example that if there is an office building in the middle of a forest, it is assumed that the people within that mobile network cell are probably mostly in the office building and not wondering in the forest. Aerial methods, such as mapping geographical attributes like road proximity, slope, land cover and nighttime lights by remote sensing are widely used in the balanced user assignment in population distribution research (Dobson et al., 2000; Ruther et al., 2015). Referring to the results of this thesis work, much simpler methods, such as just weighing lakes and forests differently than urban areas are proven to provide reasonably solid results in the balanced user assignment.

Lastly, we'll go through dasymetric user assignment. Dasymetric interpolation is a method similar to balanced user assignment but goes even deeper into different data sources utilized for estimating population distribution. As with balanced user assignment land-use types are used to estimate population distribution, the dasymetric method additionally includes datasets like mailing information (Langford et al. 2008), business addresses (Greger 2015), road proximity and traffic census results (Smith et al. 2014), POI-locations (point-of-interest) (Bakillah et al. 2014), detailed cadastral maps (Maantay et al. 2007) and building functions (Aubrecht et al. 2009, Greger 2015, Biljecki et al. 2016). The aim is to model human activities and mobility within space and time more accurately than just basing the estimations on land-use or area sizes, See Fig. 4. With this kind of comprehensive approach and good feasibility of the method, the dasymetric mapping can be said to be one of the best techniques for estimating population distribution in a given area (Mennis & Hultgren 2006; Wu et al., 2005). Research group of University of Helsinki introduced a multi-temporal function-based dasymetric interpolation method for mobile phone datasets in 2017 (Järv et al., 2017).

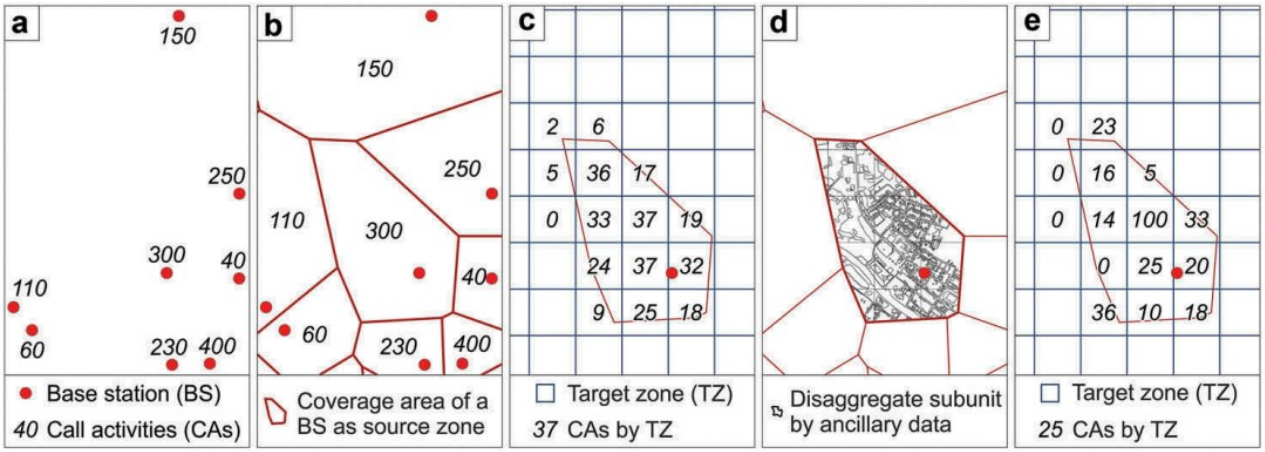


Figure 4. After the simple areal weighing method is done and the observations are distributed into the grids in part c), the dasymetric data sources are added in part (d). After dasymetric interpolation in part e), the records are distributed based on temporal human activities, land-use functions and building types, and the record counts are changed from the area proportional assignment (Järv et al., 2017).

2.7 Trip and activity assignment

So far, we have gone through how mobile phone network connections can be located geographically and how these observations can be interpolated into smaller spatial units. However, to be able to model human mobility, this huge pile of spatially distributed network events must be processed into chronologically ordered trips and activities executed by individual mobile phone users. A trip stands for a movement between two activities, whereas an activity means a person staying still performing tasks, such as being at home or doing grocery shopping at a mall.

As the user IDs and timestamps are automatically generated by the network event log files, distinguishing individual user locations in given times is easy. In order to analyze human travel behavior, these points in space and time must be then connected with each other and classified into different combinations. The most important part in this process is to categorize each of these points into one of two categories: staying point or moving point (Asakura & Hato, 2004). Once every individual location detection is classified either into moving point or staying point and required trip and activity definitions are given, it becomes possible to determine additional attributes of travel behavior such as origin, destination, home, work and so on.

Analyzing the records regarding trips and activities is also needed for possible noise detections. This happens when a mobile phone changes its connection to a different base station even though the user stays at one point. This is caused for example by fluctuation of signal strength (Asakura & Hato, 2004). Objective of these kinds of algorithms is to detect and filter out impossible user movements, such as travelling 10 kilometers in two seconds and then returning back in three seconds.

Widhalm et al., 2015 suggests a method for detecting stays and movement from mobile network data by combining properties of a low-pass filtering with an incremental clustering algorithm, basing on the work of Hariharan and Toyama, (2004). Simplified, the algorithm

clusters consecutive network events that has occurred in close proximity of each other and assumes the longer distances and time gaps between the clusters to be moving between activities, see Fig 5. The positioning errors caused by signal noise are cut out simply with a travel speed constraint.

Defining start times and end times for activities (such as arriving and leaving a shopping center) is difficult due the mobile network events occurring sparsely in time. For example, the first signaling event might occur in the shopping center before the user has left the place and the second one when the user has already been home for a while. However, method proposed by Widhalm et al., 2015 tries to estimate the departure and arrival times of the virtual locations (network event clusters) based on the average travel speeds and the distance between the locations. The stay duration can be this way bounded between the first possible arriving time and the last possible departure time. Movement directions are also considered when defining activities and their potential durations; a network event along a way of movement is much more likely to be a moving point than a network event location in which user clearly arrives from north and then leaves back again to north. Similar methods have also been proposed for example by Alexander et., al (2015), Jiang et al., (2013), Asakura & Hato, (2004), and different mobile phone data filtering techniques are being compared in the work of Horn et al., (2014).

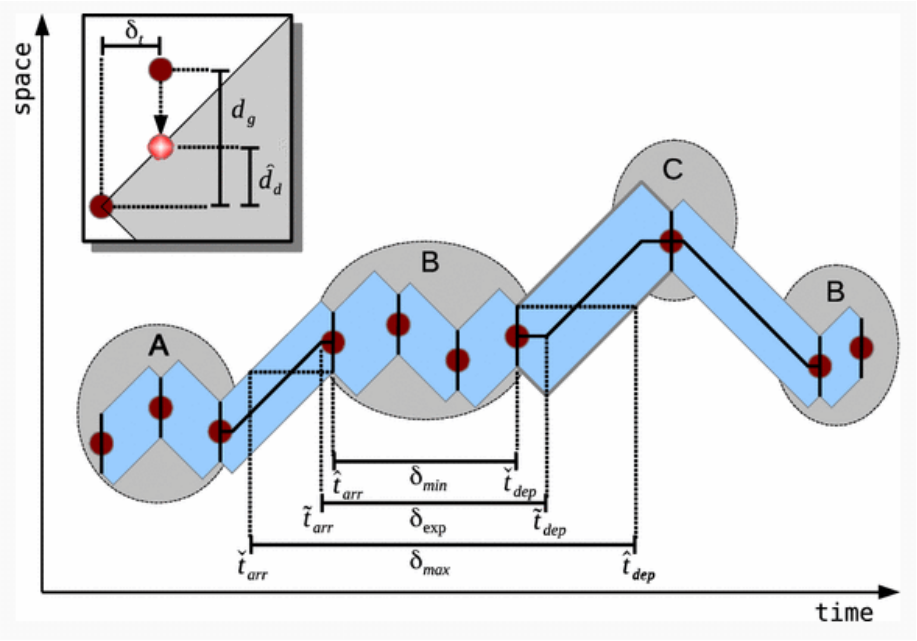


Figure 5. Network events are presented as red dots in space-time. Network event clusters are assumed to be activities and the distances between the clusters are assumed to be trips. The arrival times and departure times of the activities are estimated by an average travel speed against distances between the clusters (Widhalm, et al., 2015)

After the discrimination of trips and activities from the raw mobile phone data it has also been demonstrated that it is possible to label certain types of trips and activities into trip and activity types. (Alexander et., al 2015). As we know that human travel behavior is characterized in a way that people tend to visit same locations frequently throughout the week (Song et al., 2010b; Song et al., 2010a), it is possible to make reliable predictions about these locations based on a few attributes.

Study by Alexander et al., 2015 distinguished home and work locations from mobile phone data and this way derived trip types for home-based work (HBW), home-based other (HBO) and non-home based (NHB) trips. The home location of each user was identified by detecting the locations that have the most mobile phone network events on weekends and on weekdays during the timeframe of 7 pm and 8 am. This method is placed on the assumption that people tend to spend nights at home sleeping, or at least spend most time of the night at home. Then again work locations are identified by detecting the locations which are the furthest frequently visited locations from the home locations (max (distance from home * total number of visits). This assumption is based on the works of Levinson & Kumar, (1994) which state that working trips are usually the longest trips that people tend to make on regular basis. In case of user IDs of the phone users are not mixed in short intervals and same users can be traced longer time periods, longitudinal historical data over several months can be also used for detecting frequently visited locations and this way recognize important places of the users (Isaacman et al., 2011).

By determining and labeling trips and activities of a mobile phone user, a trajectory of the user's daily trips and activities can be estimated, see Fig 6. Regarding the type of mobile phone data and the locating techniques used, the accuracy of the trajectory can vary. In addition to detecting work and home locations (Alexander et al., 2015), land-use types, such as shopping and leisure, can be used to predict clear trip attractors (Widhalm et al., 2015) and this way activity purposes of people in different locations.

Map 1. Individual daily user trajectory. Base station locations that have provided the network signal for the user throughout the day are presented as red dots (In this case base station locations are used instead of cellular boundaries for the location estimation). The Green line is the estimated user trajectory derived from the base station locations and yellow stars represent the activities in which the user has spent time and stayed still. The numbers indicate the time of the day in hours since midnight (Widhalm et al., 2015).

2.8 Origin-Destination Matrix

Trip distribution is a factor that describes where trips originate and where they end (Ortuzar, 2011., s.175). This information is being used in transport modeling in order to support decision making and transportation planning. The idea of trip distribution is to understand how much people are moving between each location pair and this way to understand where people move and why, and which parts of the transportation network need development. In other words, trip distribution answers the question of does the transport network supply correspond with the network demand. The following step after trip distribution is usually to get a better idea of the modes used and the routes chosen on the trips between origins and destinations.

The most used way of representing trip distribution is called an Origin-Destination Matrix (ODM) (Ortuzar, 2011., s.175). ODM stores the trips made from an Origin (rows) to a Destination (columns) during a specific time period. Its idea is to present the counts as aggregated sums, not specify individuals. Additionally, modes used or activities executed can be disaggregated into the matrix. By counting the sum of trips originating from each zone and ending in each zone, each location can be classified as a trip generating or as a trip attracting zone. In a given time, trip generating zone generates more trips than receives, like residential areas at mornings. Then again trip attracting zones receive more trips than they generate, like office districts at mornings. Example of a generally formed ODM with trip counts (Table 3.):

Table 2. Example of Origin-Destination Matrix with trip counts between locations, where different locations are presented with 1, 2, 3, j and z, T_{ij} are the trips from i to j, O_i is the total number of trips originating at zone i and D_i is the total number of trips ending at zone i.

	Destinations					
Origins	**1**	**2**	**3**	**...j**	**...z**	$\sum_i T_{ij}$
1	T_{11}	T_{12}	T_{13}	$\dots T_{1j}$	$\dots T_{1z}$	O_1
2	T_{21}	T_{22}	T_{23}	$\dots T_{2j}$	$\dots T_{2z}$	O_2
3	T_{31}	T_{32}	T_{33}	$\dots T_{3j}$	$\dots T_{3z}$	O_3
⋮						
I	T_{i1}	T_{i2}	T_{i3}	$\dots T_{ij}$	$\dots T_{iz}$	O_i
⋮						
Z	T_{z1}	T_{z2}	T_{z3}	$\dots T_{zj}$	$\dots T_{zz}$	O_z
$\sum_i T_{ij}$	D_1	D_2	D_3	$\dots D_j$	$\dots D_z$	$\sum_{ij} T_{ij} = T$

Conventionally travel surveys have been the most used data source for trip distribution estimation (Alexander et al., 2015). Throughout the last decades there has been multiple different methods to analyze travel survey data and process it into trip distribution models. For example, Heanue & Pyers (1966) compared following trip distribution procedures in their study: a growth factor model, a gravity model, an intervening opportunities model and

a competing opportunity model, which all of them base on mathematic estimations about people's movements. However, in modern context these trip distribution methods based on survey data "are not able to keep the pace of city growth and change, making relevant dynamic phenomena to be invisible for transportation and urban planners" (Graells-Garrido, 2018).

The usage of mobile network data to build Origin-Destination matrices was first introduced in Italy by Bolla & Davoli (2000) and tested with small samples on specific roads by White & Wells (2002). At the same time White and Wells (2002) tested also constructing an ODM in England from mobile phone billing data, which turned out to be not accurate enough for the purpose. After that, dozens of studies have demonstrated that mobile network data can be used to produce OD-matrices (e.g. Alexander et al., 2015; Bonnel et al., 2018; Iqbal et al., 2014). The main differences between mobile network data and survey data regarding OD-matrices are roughly the same that the differences between these datasets in overall, which are presented in the section 2.12. However, an aspect which makes a clear difference between these two is that mobile network data can be processed into an ODM in a reasonably realistic way, without calibration methods such as opportunity model, gravity model, maximum likelihood or optimization (Alexander et al., 2015). Study by Alexander et al., (2015) demonstrates also that the zonal aggregation sizes of origins and destinations may have an impact on the accuracy of ODM derived from mobile network data compared to ODM derived from surveys. Their results show high coherency between the datasets in city-to-city OD-matrices, whereas in the comparison of zip-to-zip OD-matrices the differences are larger. One proposed reason for this was that the mobile network data did not have the spatial accuracy required for the small-scale trips within cities. In overall, their OD comparisons between the datasets showed high coherency if larger than 2,5 square kilometer zonal aggregations were used. However, mobile network spatial accuracies can vary significantly between different mobile network technologies and between the types and methods of mobile network data gathered (see sections 2.4 and 2.5), which affect directly to the reliable aggregations sizes that can be used in OD-matrices derived from mobile network data.

Another aspect that creates differences between the results of algorithm-based and survey-based OD-matrices is the definition of a trip. Whereas mobile network data algorithms simply define an activity based on the time that is spent still, surveys might allow surprisingly long breaks during the trips, if the performed break is "part of the trip". This phenomenon is discussed more further in the thesis (see sections 3.1.2 and 5.4.2).

2.9 Representativeness

One major challenge in human mobility studies regarding the usage of mobile network data is that usually the field of mobile phone subscriptions is highly competed by several mobile network operators. This means that it is extremely difficult to acquire access to mobile network events of the whole population (100 % of the residents in an area) because the population is distributed between several operators. Actually, no study before 2018 had been reported being made with a mobile network dataset which would have been statistically suitable to present the whole population (Horanont et al., 2018). As users of mobile network data are not usually interested in the mobility of one operator's subscribers but in the mobility of the whole population, questions rise whether the subscriptions of one operator

can represent the whole population, what are the factors affecting this and is it possible to extrapolate movements of whole population from datasets of one operator.

Teerayt Horanont, Thananut Philboonbanakit and Santi Phithakkitnukoon conducted a study about representativeness of CDR records in the fall of 2018. At the time when this thesis work is being written, this study remains as the first and only work which has studied mobile network data representativeness regarding operator biases. Horanont et al., 2018 examined the effects of three aspects of mobile network datasets by different operators regarding the representativeness of the data; operator's market share, the urban-rural user population ratio and user gender ratio.

The study by Horanont et al., 2018 was made in Asian city with 1,6 million residents served by five mobile network operators. Market shares of the operators were as follows: Operator 1: 65,31 %, Operator 2: 30,27 %, Operator 3: 4,22 %, Operator 4: 0,1 % and Operator 5: 0,1 %. The correlation between the population density values by census and CDR data from each operator with a different market share was examined. The results of the correlations between each operator against the actual population density by census are presented in Table 2. below:

Table 3. The R-squared correlation values between CDR-based population density and census-based population density by five different operators in Asian city (Horanont et al., 2018)

Operator	Market share (%)	R-squared
1	65.31	0.88
2	30.27	0.82
3	4.22	0.80
4	0.10	0.70
5	0.10	0.76

As we can see, higher operator market shares correlate with the census values better than the lower market shares, but the differences are surprisingly low. Increase in the market share is not directly proportional to the increase in the correlation value, as even small market share operators are able to provide reasonably high correlation values compared to large operators. In other words, size of the market share does matter, but also urban-rural ratio and gender ratio play important roles in the representativeness of mobile network dataset.

When it comes to the urban-rural ratio and the gender ratio, Horanont et al., 2018 examined in overall what kind of ratios should an operator have to best represent the whole population. By comparing CDR record datasets with different distributions of urban home locations, rural home locations, male network events and female network events to the census values, they were able to distinguish the datasets that matched the census values best. Study results stated that an operator with user distribution of 80 % living in urban areas and 20 % in rural areas represents the whole population best. The best matching gender ratio was 50-50. In other words, the study proposed that even smaller market share operators can provide reasonably representative datasets if their gender ratio and geographical ratio are balanced accordingly. However, the spatial coverage of the study provided view only into one city,

and this way the results are not necessarily applicable to larger spatial extents. In addition, Asian mobile network behavior and mobility behavior might not be directly applicable in Nordics. Nonetheless, study by Horanont et al., 2018 gives a solid view into fundamentals of mobile network representativeness and into factors that operators should consider when extrapolating their records to represent the whole populations.

When it comes to extrapolation, information like mobile network subscriber's home location or gender are not necessarily available due the privacy issues (Bonnel et al., 2018) and referring to study by Horanont et al., 2018, this creates challenges regarding credible extrapolation. Study by Bonnel et al., 2018 compared results of a travel survey into a mobile network dataset similarly to this thesis work in 2017. Their mobile network operator had similarly to Telia roughly a market share of 33 % and they used a simple extrapolation method which based on the number of mobile network subscribers' network events per administrative area and the number of people living on the area, as not deeper knowledge of the subscribers was available. The following formula was used:

$$C_{exp} = \frac{Population\ of\ RA\ region}{Nb\ of\ users\ using\ 3G\ network}$$

Where C_{exp} is the expansion factor, "population of RA region" is the population of specific administrative area inferred from census data, and "Nb of users using 3G network" is the number of observed telephones with at least 4 events occurred during a day (Bonnel et al., 2018). Minimum of 4 network events was set due the aim of the study which was to produce Origin-Destination matrix. Basically, their extrapolation method was based only on the market share of a single operator. However, reasonably good results were still generated.

2.10 Privacy

The usage and development of location-based services have raised a lot of discussion regarding the privacy policies of the services. Modern technologies such as internet and mobile devices have made it possible for the service providers to accurately track down individuals (De Montjoye et al., 2013), and use this information to generate business. Even if the datasets containing personal information were used responsibly by the operators and by the potential third parties, they may be exposed to cyber-attacks and be utilized in criminal purposes (Gedik & Liu, 2008). Hence, the sharing and gathering of mobile phone data raises serious privacy issues as the network connection events of the users can reveal private aspects of individuals' travel behavior and life (Calabrese et al., 2015). Study by De Montjoye et al., (2013) states that with a database of network events generated with a time interval of one hour and spatial accuracy of a Cell ID, it is possible to uniquely identify 95 % of the individuals with just four spatio-temporal points. The Uniqueness of human travel behavior is therefore actually very high and things that can be deduced from the visited locations range wide. This is the reason why effective anonymization techniques must be implemented to mobile network data systems.

In addition to regulatory strategies executed by governments and trust-based privacy policy agreements between parties, Krumm et al., 2009 outlines two computational techniques for ensuring mobile phone user privacy in location-based services. The first one is anonymization, which includes several methods on how to make the mobile phone users anonymous and this way "hide" the personal trajectories. The second one bases on blurring

the accurate locations of the users, making it hard for one to detect individual locations accurately and this way deduce the personal information.

Probably the most obvious way of ensuring the privacy of location data is to *anonymize* the users. This can be done simply for example by handling users only with untraceable user ID's instead of names (Krumm et al., 2009). As untraceable user ID's might not actually in real life be that untraceable, the user ID's can be shuffled (Widhalm et al., 2015) frequently in order to make it harder for the attacker to access enough history on a victim for identification (Beresford & Stajano, 2003). This of course makes it harder for the data analyst to derive home or work locations of the user (See section 2.7).

By only replacing names of the users with ID's it remains possible to identify a person by joining other data sources into one that has been anonymized. A widely used method to take the anonymization into next level is called k-Anonymity (Gedik & Liu, 2008). An observation is considered k-anonymous only if k-(amount) of identical observations are detected. This means that if k-5 value is determined for an Origin-Destination matrix, the results between an origin and a destination are given only if five or more identical trips have occurred between those locations. This applies also for persons staying still; it would not then be possible in this case to detect individual persons living or moving in a forest if there are no at least four other people in the same mobile network cell. Study by Gedik & Liu, (2008) states that k-anonymity is an effective method for supporting location anonymity and privacy related to mobile network data. In addition, study by Kido et al., (2005) suggest a technique for implementing false locations into the dataset, making it possible to query the valid locations only by the users that are allowed to use the information.

The privacy of mobile phone users can be enhanced also by degrading the quality of the location measurements (Duckham & Kulik, 2005). *Obfuscation* of the location means either inaccuracy of the measurement or imprecision of the measurement. These kinds of methods are utilized especially in work with GPS-locators, in which the location information is much more accurate than with cell-based mobile network data (Krumm et al., 2009). Similarly to Kido et al., (2005), study by Mir et al. (2013), proposed a method to generate synthetic Call Detail Data records with synthetic locations and associated times to obtain the mobility patterns of large populations while preserving privacy.

2.11 Conventional Transport Data

Conventional transport data gathering has based for decades mostly on few methods: travel surveys, traffic census, roadside interviews and focus groups (Calabrese et a., 2013; Stewart & Shamdasani, 2014; Tolouei et al., 2017). This section opens basics of the conventional transport data sources with examples from United States and Finland. The major differences, advantages and weaknesses between conventional transport data and mobile network data in transport studies are presented in the next section (see section 2.12).

National Household Travel Survey (NHTS) is the authoritative source on the travel behavior of the American public (U.S. Department of Transportation, 2017). It is the only nation-wide dataset which includes travel behavior data on household level and on individual level. The data is gathered by survey with a list of questions and sent to some of the Americans in about every eight years. In the NTHS 2017, there were answers included from approximately 130 000 households.

NTHS includes data about vehicles, vehicle trips, vehicle miles, persons, person trips, person miles, households, workers and drivers (U.S. Department of Transportation, 2017). It is designed to give as comprehensive picture of the answerer as possible, so that the transportation planners could not only utilize the trips made by the answerer, but also what kind of a person the answerer is (Income level, household size etc.). This data is gathered for example with questions regarding trip purpose, travel mode, travel length etc.

National Travel Survey (NTS) by Finnish Transport Agency is a Finnish counterpart for NHTS in the U.S. and represents the basis of the Finnish travel (Finnish Transport Agency, 2018a). For the first time in 2016, the data was gathered through three different survey methods; phone call, internet, and mail. Finnish municipalities have supplemented the nation-wide results with their own local survey result datasets, enabling the total amount of 30 000 answerers for the survey. The main objective of the survey is to enable assessing of trip counts, travel times and travel distances, compared to the travel habits of Finns and the demographic, geographic and temporal variations in travel. The survey is repeated in six-year intervals.

There is a large amount of background information gathered from the Finnish answerers: age, gender, type of accommodation, size of the household, occupation, education, driver's license, vehicles of the household, option for using a car, option for carpooling, and the address of home, school or work place and potential cottage (Finnish Transport Agency, 2018a). Reporting related to transportation included the number of trips made, timing of the trips, origins and destinations, the mode used and potential use of another mode. In addition, the locations of the answerers' home and workplace were enriched with statistical data about the areas.

Also, the national travel survey dataset in Finland includes added information about how the trips between answerers' origins and destinations would have been optimally made with public transit or by a personal motorized vehicle, these analyses basing on other service provider datasets, like HERE maps (Finnish Transport Agency, 2018b). This makes it possible to assess relationship between transportation and land use, and especially assess the relationship between transportation system and individual's travel choices. In other words; why did this person choose to travel to work by train, even though a bus would have been a faster choice.

Travel surveys like NTHS and NTS can be supplemented with traffic census data (Finnish Transport Agency, 2018a). Traffic census data in Finland is mostly produced by Finnish Transport Agency. Traffic flow monitoring is made with Transport Monitoring System (TMS) (Liikenteen automaattinen mittausjärjestelmä, LAM) (Finnish Transport Agency, 2016). In 2016 there were 460 TMS points in Finland, situated to gather as comprehensive picture of the Finnish main road travel as possible. Main roads and highways are usually monitored with LAM, whereas smaller roads are measured manually when needed. In addition to monitoring traffic flow, TMS system is able to identify a type and speed of the passing vehicle. Also other additional methods, such as automatic number plate recognition, video survey junction counts and roadside interviews can be mentioned as conventional transport data services on the market (TrafficWatch, 2019), used on the side of travel surveys.

2.12 Comparison Between Conventional Transport Data and Mobile Network Data

When put side by side, conventional transport data sources (mostly travel surveys & traffic census) and mobile network datasets have a set of significant differences, which include both advantages and disadvantages regarding both data sources. However, the data truthfulness is not one of the differences; it has been demonstrated with a large number of studies that mobile network data can be used to deduce high quality OD-matrices and individual trip trajectories that are coherent with household surveys (e.g. Alexander et al., 2015; Calabrese et al., 2011; Iqbal et al., 2014; Jiang et al., 2013; Schneider et al., 2013; Zhang et al., 2010). Bonnel et al., (2018) actually state that travel surveys are "much less useful" for constructing Origin-Destination Matrix than mobile network data. Hence, the difference between datasets comes from the data characteristics like utilization costs, temporal attributes, background information and sample sizes (Alexander et al., 2015). These differences are being presented next.

One of the key advantages of mobile network data is its temporality (Widhalm et al., 2015). As every single network connection produces a timestamp, mobile devices can be located all the time, everywhere, and all around the year. In addition, the data is gathered constantly, making it possible to query datasets with specific time period and high temporal accuracy. This could be for example a location of every device by one-hour interval from the summer holiday peak of 2018. When it comes to travel surveys, the datasets represent a static moment in time and is being updated only every 5 to 10 years (U.S. Department of Transportation, 2017). This makes it difficult to use the survey data to asses travel differences in different time periods, for example during winters, summers, holiday seasons, rainy days, cold days or bus company strike days. The transport system also tends to change and evolve over time, which is problematic for low frequency travel survey production. As mobile network data is able to provide temporal attributes that are not reachable through surveys, it can supplement the travel survey datasets by giving wider view of the seasonal changes in travel (Alexander et al., 2015). It might also be worth of noticing that as mobile network data might be more capable of capturing evening and night time movement than surveys (Alexander et al., 2015), it might also be more capable of capturing movements in rural and low-density areas that are not monitored that well or at all (Bhekor et al., 2018; Cheng et al., 2006; Demissie et al., 2013).

When travel surveys tend to base on regularity, mobile network data is able to show anomalies in travel (Calabrese et al., 2015). This means, that due the mobile network data temporal attributes, it is possible to see how unusual or accidental events may impact the transportation system. This could mean for example concerts, events, accidents, road blocks, strikes or storms. Travel surveys do not have this kind of an ability, as their questions are usually formed in a way which gathers regular information of the answerers daily travel (Alexander et al., 2015).

Third major advantage of mobile network data is its low cost of production (Alexander et al., 2015). Finnish National Travel Survey 2016 employed directly or indirectly hundreds of people and was gathered and enriched by several private companies and dozens of public organizations (Finnish Transport Agency, 2018a). Compared to mobile network data, which is constantly being gathered whether or not it is used for transport planning, travel surveys are remarkably more expensive to produce. It is of course a different story whether telecom operators make the data available and with what price.

Sample sizes are usually also significantly different between travel surveys and mobile network datasets (Tolouei et al., 2015; Tolouei et al., 2017). Referring to Horanont et al., (2018), regular customer share between operators in Asia goes roughly as follows; one large operator (about 50 % market share), few medium sized operators (about 20 % market share) and handful of small operators with a few percent market share. Similar results can be applied into Nordics, where the largest operator goes usually with a market share of 30 – 40 % (Traficom, 2017). This means that the largest operators are always aware of movements of massive amounts of people, as the travel surveys cover only tiny fractions of the populations (e.g. National Travel Survey of Finland 2016: coverage of 0,5 %) (Finnish Transport Agency, 2018a). This means that also geographical coverage is usually more representative with mobile network datasets than in travel surveys (Tolouei et al., 2015). Small sample sizes are problematic also in a way that it gets difficult to distinguish specific elements from the data. For example, in long distance traveling one cannot distinguish a travel direction between origin and destination with Finnish National Travel Survey, as that would make the sample size too small. Only the sum of the trips setting to both directions are used (Finnish Transport Agency, 2018a, s.107).

However, survey datasets have their own benefits when compared to mobile network data. One of the biggest disadvantages of mobile network data against survey/census data is its lack of background information of the mobile device carrier (Alexander et al., 2015). As a survey is able to gather basically any information desired, background information of the mobile network event is significantly more limited. Even though mobile network data can be enriched with statistical information of the user's theoretical home and workplace locations, and the mobile operator may be aware of the gender and the age of the customer, it cannot easily sort out whether the user is a student or a retiree, rich or poor, or how the user lives. Also other trip attributes, such as mode or purpose, are significantly more difficult to derive from mobile network data than with surveys.

Data accuracy must also be considered when comparing conventional transport data and mobile network data. As trips and activities derived from mobile network data are always defined algorithmically, their accuracy may vary between different parameters, different datasets and data providers (Tolouei, 2015). In large scale analysis it can be crucial whether activities are distinguished from trips with 30-minute stopping time or with 40-minute stopping time (Bonnel et al., 2018). Then again, also travel survey and roadside survey result accuracies can be questioned. As it can be seen from the results of Finnish National Travel Survey 2016, people might not be highly accurate with their answers. For example, it is a common phenomenon that people round up or round down the length of their long-distance trip travel durations, instead of accurately submitting the actual travel times. In addition, it is demonstrated that number of trips are usually being unreported and underestimated in household travel surveys (Wolf et al., 2003).

Global Positioning System (GPS) has been probably the most significant technology used to overcome the deficiencies of travel surveys and census after late 1990s (Huang et al., 2019). The most remarkable advantage of GPS-data compared to mobile network data is its temporal and spatial accuracy. Basically, GPS provides uninterrupted movement trace with spatial accuracy of only few meters. However, it faces problems in study durations and scales, as GPS loggers need to be carried with the participants or be actively enabled in smartphones. GPS-data sample sizes are this way relatively small, and data inconsistencies (GPS is turned off during the study) are regular (Wu et al., 2016). Hence, it is difficult to conduct large origin-destination matrices from GPS data. In addition, participants of the

National Travel Survey 2016 stated that they were not willing to keep on GPS during the study period for privacy reasons (Finnish Transport Agency, 2018b).

2.13 Modal Split

The choice of transport mode has been probably one of the most important classic model phases in transportation planning (Ortuzar, 2011, s. 207). Reason for this is that public transport plays a key role in public policy making due its cost-effectiveness, and the only way to examine its utilization is to understand how people travel between locations and why they choose the mode they end up using. In addition, the importance of sustainability has had an accelerating role in the modal split modeling due the negative externalities of the usage of private motorized vehicles. By referring to Ortuzar (2011), conventional modal split modeling bases on aggregated or disaggregated survey data, which is then processed and extrapolated with mathematical models, such as binary logit models, hierarchical modal-split models and direct demand models. All of these methods try to estimate the most cost-effective mode for a given OD-pair and this way assign portions of people into each mode.

During last few years while network-event mobile network datasets have emerged more and more, potential of inferring modal split from mobile network data has been understood and started to study. Huang et al. (2019) searched and listed all peer-reviewed scientific studies about mobile network data modal split in the April of 2019 and demonstrated that there are 22 studies regarding modal split from mobile network data, from which almost all of them are published during the last five years. The number of modes detected ranged mostly from two to three, and many of the studies were concentrating only on intercity trips as the modes are easier to distinguish from longer trips. However, study by Danafar et al. (2017) tried to differentiate as many as six modes; car, bus, tram, train, cycling and walking. 16 out of 22 studies used geographic data joined with the mobile network event locations, as vehicles tend to travel on known links between known stations and this information can be utilized in the mode detection algorithms. In addition, Yamada et al. (2016) and Horn et al (2017) considered train timetables to improve the performance of their modal split algorithms.

The most utilized method in modal detection has been the rule-base heuristics (Huang et al., 2019). This means that simple rules which determine the mode are given for the data. For example, if the travel speed is over 300 km / h, an airplane is assigned, whereas if the travel speed is under 300 km / h and a road network is in closer proximity of the trip than a rail network, a car is assigned. In addition, Qu et al., (2015) used more complicated rules for the detection, such as proximity to public transport network and a logit model. 18 of the total 22 studies used a rule which considered a proximity to either road network or to public transport network, which demonstrates the high utilization of the rule-based detection combined to additional transportation GIS-data.

In addition to rule-based detection, a handful of methods have been developed for assigning the trips that are left unassigned by the rules (e. g. Kalatian and Shafahi, 2016; Wang et al., 2010; Xu et al., 2011). These unassigned trips are usually characterized by attributes that do not directly promote any of the modes considered by the rules but may still include some hints regarding the most probable modes used. Basically, this clustering of unassigned trips can be made manually using common sense or automatically using unsupervised machine learning (Jahangiri et al., 2015). However, referring to Huang et al., (2019), not enough validation has been performed for the different methods to draw conclusions regarding which method works best.

3 Data & Data Processing

Aim of this chapter is to present data sources that were used to validate Telia's mobile network data regarding long distance trips, route choices, travel durations and modal split. This validation was done to demonstrate that mobile network data can provide reliable information about people's movements and additionally give insights into aspects that cannot be derived with other data sources.

So, in addition to the mobile network data itself, the thesis work project included reference data which was analyzed and compared against the mobile network data. In other words, movement attributes like trips, modal split and route choices between an origin and a destination were derived both from Telia's mobile network data and from other data sources, and then compared whether they match.

Like stated already in the Introduction -chapter, this chapter presents the elements of the thesis work project that existed already before starting the project. This means all the mobile network data processing algorithms of Telia and the reference data sources. Even though it might be little confusing to call the data processing algorithms and Telia's mobile network infrastructure as data of the project, they are chosen to be presented in the Data & Data Processing -chapter and not in the Methods -chapter, as the Methods -chapter focuses only on the things that are done explicitly in the name of the thesis work, not anything that is made by someone else before. The separation between data and methods of the study is not unambiguous, so the best way to demonstrate it is a project flow chart (see section 4.1); Data & Data Processing -chapter contains the phases 0 and 1.

The most significant reference data source of the study is the National Travel survey of 2016. The NTS is presented in full detail in this chapter including its extrapolation and gathering methodology. In addition, LAM sensors, HELMET transport demand model and Finavia flight statistics are used in the validation. These reference data sources are introduced shortly in the third section of this chapter called "Other Reference Data Sources".

The structure of this chapter is as follows: First, the whole Telia's data production pipeline is presented. Then, we move to the National Travel Survey 2016 and finally to the other reference data sources.

3.1 Telia's Mobile Network Data Infrastructure

This section demonstrates how the mobile network data is handled by Telia and how the trip chains and OD-matrices are created from the actual signaling events received from the moving mobile phones. This complicated set of processes and algorithms is too wide to be explained in full detail, so only the key elements are explained. In addition, detailed information about algorithms and subcontractors used in the process classify as trade secrets and cannot be shared publicly. However, a high-profile demonstration provides a clear enough picture about the data processes used in the study. The section is divided into sub-sections representing each phase of Telia's mobile network data processing pipeline (Figure 6.).

The following sub-sections about Telia's mobile network data infrastructures and processes follow the same themes that are presented in the Theory-chapter. Whereas the Theory-chapter describes the general terminology and other studies related to these methods and

processes, this chapter specifies how Telia handles its data productioning and what kind of technology and methodology it uses in its data processing. The parameters that were used in the generation of mobile network data OD-matrices are given in chapter 4., Methods.

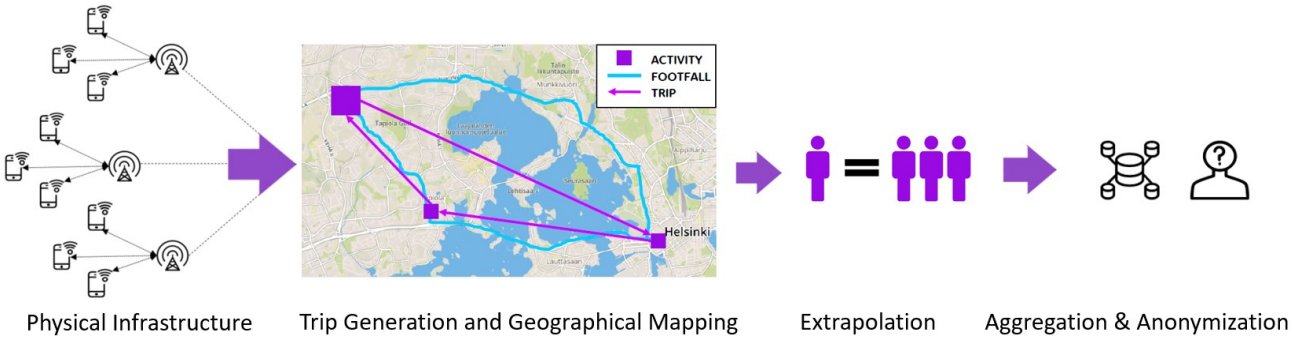


Figure 6. Telia's mobile network data processing pipeline.

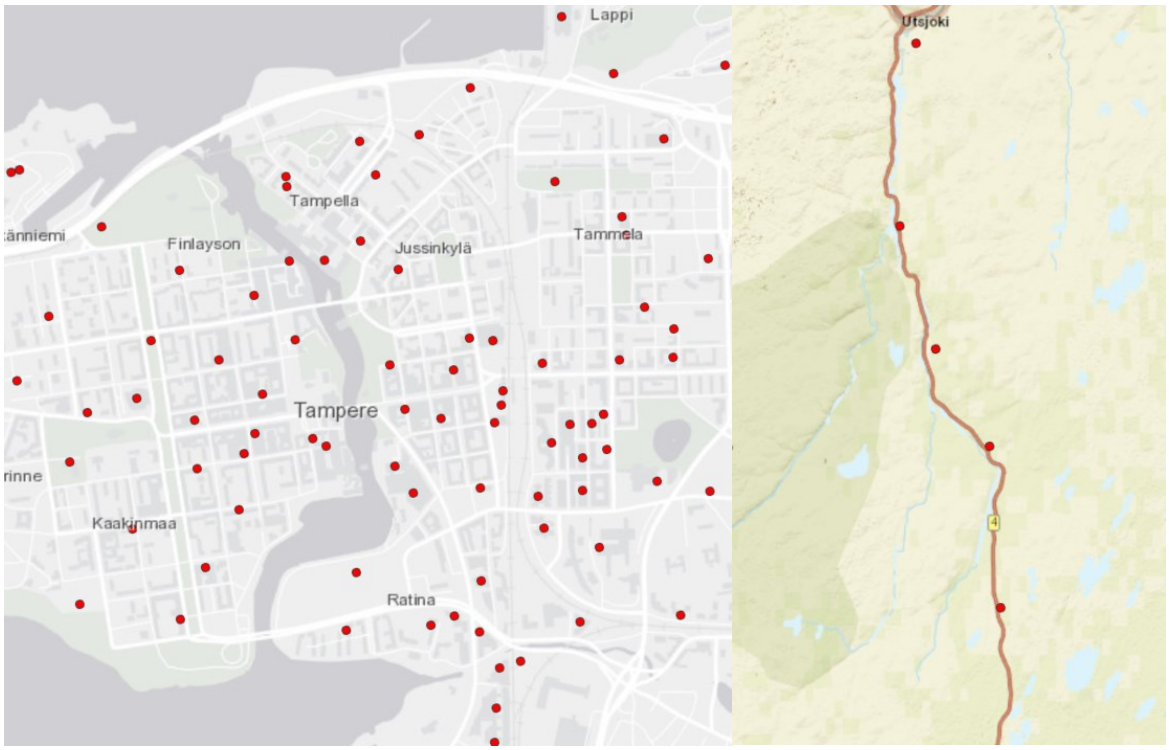
3.1.1 Physical infrastructure

To generate its mobile network, Telia uses roughly 50 000 base stations in Finland, in which it has roughly 100 000 antennas. There are different antennas for different kind of network types (2G, 3G, 4G and 5G), and all the base stations do not necessarily contain antennas with all types. However, the technical differences between antenna types do not make a difference in mobility insights generated from mobile network data.

The distances between base stations can be classified to three classes. In cities, base stations are typically 250 – 500 meters from each other (Picture 1.), whereas in rural areas the distances between base stations are longer, roughly 5 – 7 km. The third class, wilderness of Lapland, contains yet smaller amount of base stations; there the typical distance between base stations is from 15 km to 30 km or somewhere even more.

As Telia uses CGI (Cell Global Identity, see section 2.5) as its mobile network positioning method, the distance between base stations plays a significant role. The positioning accuracy is directly proportional to the mobile network cell sizes, which are then again directly proportional to the distances between nearest base stations. The positioning bases on BSEs (Best Server Estimate), which means that if a person is connected to a certain base station, it is assumed that the person is currently located in the area which is best served by that specific base station (which is typically the nearest one).

However, the positioning of the base stations is even more important than the distances between the stations. The fact that base stations are located near roads and crossroads (Map 2.) makes it harder to locate people moving in the wilderness, but significantly easier to locate people moving on the roads. This is relevant from the viewpoint of transportation studies, as people tend to move using roads, especially if the distance travelled is long. As it might be that there are few wanderers in Lapland that are not connected to the mobile network, Telia is able to track down people moving on the roads, which accumulates the huge majority of the people moving in total.



Map 2. Base stations in the city of Tampere (left) and base stations on the road 4 in Lapland, from Inari to Utsjoki (right). Even though wilderness of Lapland is not densely served with base stations, the roads are covered comprehensively with stations between 5 – 10 kilometres.

3.1.2 Trip Generation

Telia uses network-driven mobile network data technology in its mobile network data gathering (For event-driven and network-driven mobile network data, see section 2.4). With passive and active mobile network signaling events combined, Telia is able to conduct in average 500 – 1000 network events from a single user during one day. For a regular day time activity, this means signaling events occurring on average between 1 – 3 minutes. Hence, movement patterns of Telia's subscribers are captured with high accuracy, and biases do not occur due too little amount of signaling events.

Each signaling event includes information regarding the location of the phone (Unique cell ID (CI-LAC)), and anonymous user ID (which however can be joined with information regarding age class and gender of the specific user ID). Unlike with widely used CDR -data, Telia's network driven signaling events do not include information regarding the type of the connection (SMS, call, call duration, etc.) or information about the possible respondent of a call or a SMS-message. Demonstration of Telia's mobile network data log for one user moving:

Table 4. Demonstration of Telia's mobile network data log. The figurative user seems to be on the move, as his or her phone is connected to different antennas over time.

User ID	CI-LAC	Time
95639	43200	14:01
95639	43301	14:02
95639	45234	14:03
95639	45701	14:04

After the raw mobile network signaling event dataset is gathered, the signaling events are classified into moving points and staying points (See section 2.6). Basically this means, that if there is a cluster of network events in close geographical proximity of each other over a long period of time, it is likely that the user has stayed still in that location performing an activity. If the network events then again seem to be progressively proceeding into some direction over time (Table 4.), the user is likely to be on the move and the signaling events are combined and classified as a trip.

The most important factor in the classification of moving points and staying points is determining the breaking factor which stops a trip and starts an activity. For example, if a person is driving from Helsinki to Oulu and stops for a lunch in Jyväskylä, the algorithm must decide whether there was a trip from Helsinki to Jyväskylä and from Jyväskylä to Oulu, or only one trip from Helsinki to Oulu. This becomes crucial especially when conducting long distance OD-matrices, which aim to show where people truly start and end their trips, not where they stop on the way by.

In Telia's algorithms the classification of trips and activities are handled with break parameters, which makes it possible to manually iterate the effect of different break parameters to the final outcome. There are two factors which affect to the ending of a trip; a break time and a directional change. If the break time is longer than the given allowed break time, the trip is ended, and an activity started. For example, if the lunch break in Jyväskylä is longer than one hour, the trip is ended. Then again if the direction of a trip changes more than the given allowed directional change, the trip is ended, and an activity started. For example, if a person drives to the airport to pick up someone and then drives immediately back home, the one constant tour is divided into two trips; a trip from home to the airport and a trip from the airport to home.

The break factor can also be, and often is, a combination of the allowed break time and the directional change. This way many kinds of situations, in addition to the examples of lunch break and airport pick up, can be handled accordingly. The allowed break time can also be set to vary distance dependently. This means that longer trips are allowed for a longer break without stopping the trip than shorter trips. For example, a one-hour lunch break is allowed for a 600 km trip from Helsinki to Oulu, but not for a 2 km long trip from home to school. In short trips the shortest breaking time is set to about 15 minutes, as this is the maximum time that it usually takes to wait a bus or a train on a public transport stop.

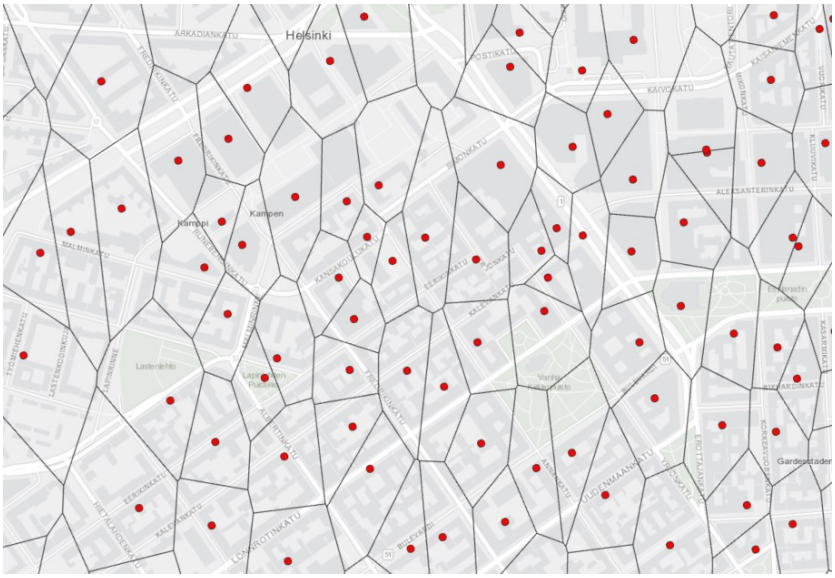
Telia's signal noise detection is handled in the same phase with the classification of trips and activities (For noise detection, see section 2.6). There are a handful of simple travel speed constraints that are used in the noise filtering, such as maximum ground velocity of 250 km/h, minimum air distance of 100 km and minimum air velocity of 400 km/h. If the trip assignment algorithm detects movement that exceeds these limits, these trips are filtered out as noise.

The most important parameters used in the trip generation process in the thesis work:

- Break parameter: Distant dependent (If distance below 10 km, maximum break 12 min. If distance over 800 km, maximum break 70 min. Between these limit points increasing break trend from the shortest break value to the longest break value.)
- Delta bearing (directional change): If the trip is under 10 km long, 15 min break is allowed instead of 12 min, if the direction of the trip changes over 20 degrees. Between the limits regarding distance dependent break value, also the break extension due the directional change is applied with an increasing trend accordingly. However, if the directional change is over 140 degrees, the trip is always ended and a new trip started.

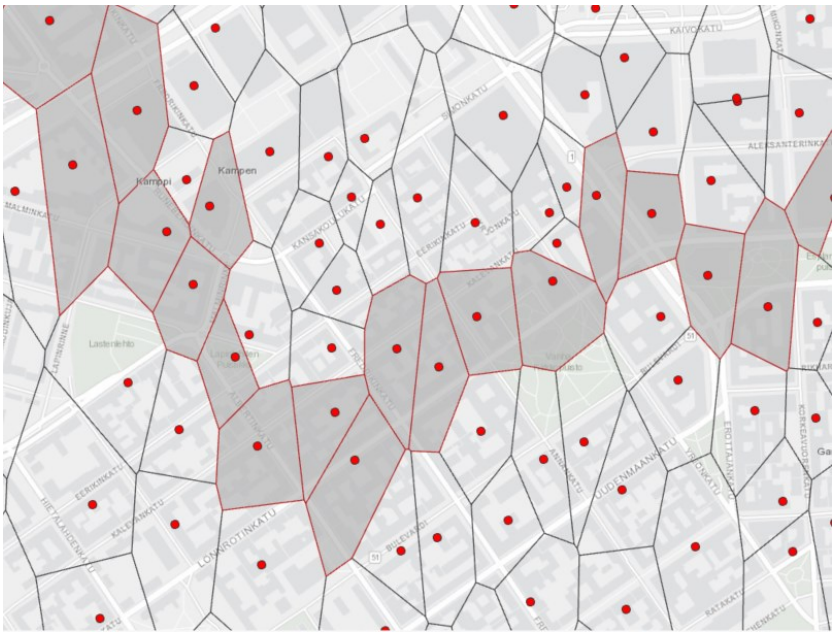
3.1.3 Geographical mapping

Once the network-driven mobile network data log is generated (Table 4.) and the signaling events are classified into trips and activities, the trip chains and activities can be mapped. This mapping bases on the fact that Telia is aware of its base station locations and Best-Server-estimates (cell boundaries):



Map 3. Illustration of base stations (red dots) in Helsinki and Best-Server-Estimates (cell boundaries) (black lines) that the base stations serve. The illustration is made with Voronoi tessellation (Gonzales et al., 2008) – actual division of BSE:s is more complicated, as base stations using different technology provide different sized cells. Also factors such as topography and buildings affect to real BSE:s.

Once a mobile phone with Telia's subscription moves through an area, it is connected to the base stations which provide the cells in which the user moves. This creates the mobile network data log (Table 4.). At this point the movement can be mapped in the name of the cell ID:s used:



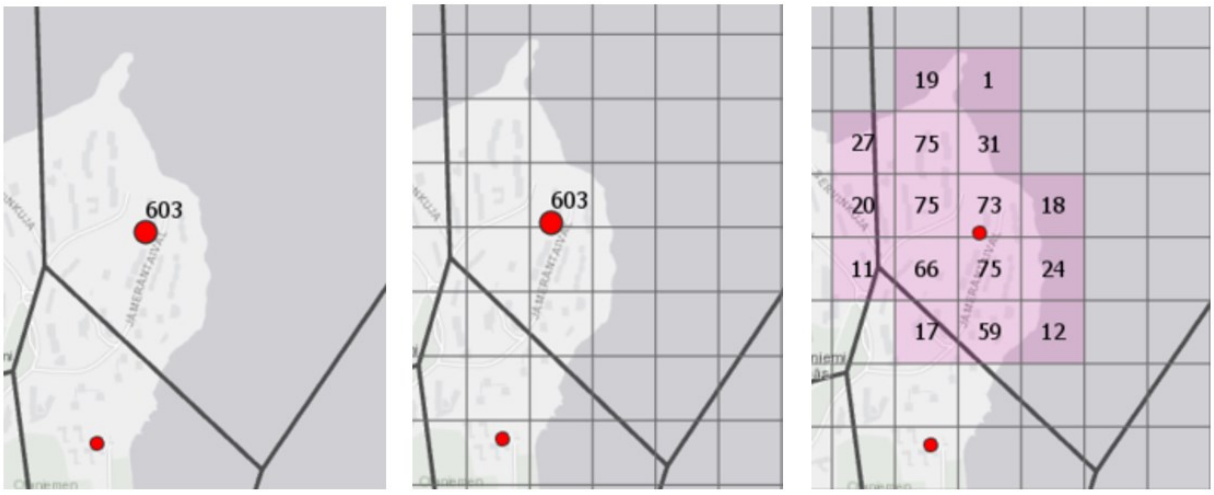
Map 4. The cells which were used by the person who travelled through central Helsinki are highlighted with darker grey.

At this point Telia has already a relatively good picture of the path that was used by the person. However, also other geographical data can be used to estimate even more accurate paths and locations. For example, origin-destination analysis assigns the "cell paths" into the Open Street Map road network to enable information on route choice (Map 5.). Open Street Map (OSM) links and their classifications are used in the assignment in a way that lower level links are weighted in the start and in the end points of a trip and higher capacity links are weighed on the core part of the trip.

Map 5. The road that the person uses when traveling through an area can be estimated (red line) even though the mobile network positioning does not provide that accurate location information.

In addition to trips, also activities are mapped. For example, if a person is connected to the same base station fifteen times during two hours, a two-hour long activity is created to that specific cell. If the person is also connected to neighboring cells during that two-hour activity, the best matching cell is chosen for the location of the activity. This is relevant especially in the areas, where there is a lot of base stations nearby; the base station which provides the connection might change during an activity, even though the person doesn't really perform a trip to another location. For example, walking from a conference room to another inside an office building might change the connection to another base station. From the viewpoint of an OD-matrix, this is relevant as the activity locations determine the origin zones and the destination zones of the matrix.

When it comes to user assignment, Telia uses balanced user assignment method (for different user assignment methods, see section 2.6). This means, that when signaling events from each cell are assigned into another spatial units (like into administrative areas, postal areas or grids (Map 6.), water areas are filtered out from the assignment. So even though a coastal mobile network cell would cover an area from the sea, the users that are connected to that base station are assumed to be on the land:



Map 6. There are 603 people connected to a certain base station during an evening time hour (left). If we want to analyze more accurately where the 603 people are located, the area is divided into grids (center). The balanced user assignment assigns the 603 people into the grids by weighing out the areas that are on the water or partly on the other cells (right).

3.1.4 Extrapolation

After the trips and activities are distinguished and assigned geographically, the results are extrapolated (for representativeness, see section 2.9). Aim of the extrapolation is to present the movements of the whole Finnish population based on the subscriptions of Telia.

Based on the publication released by The Ministry of Transport and Communications in 31.7.2018, Telia has a market share of 34 % in Finnish telecommunications (Viestintäministeriö, 2018). In 2017, there were 1,7 mobile network subscriptions in Finland per capita (Traficom 2017), which means that Telia has about three million subscriptions in Finland. However, company owned subscriptions and multiple subscriptions per person cannot be used in the extrapolation and little part of the population (young children and some older people) do not own phones, so the total subscription count is slightly less.

In the big picture, the customer share of Telia is quantitatively large enough sample size to be extrapolated for the whole population (Horanont et al., 2018). However, Telia's urban – rural ratio is not optimal based on the study by Horanont et al. (2018), as Telia has higher market share in the rural areas than in large cities and not the other way around as the study suggests. In Lapland the customer shares are roughly 50 %, in inner Finland between 30 % and 50 % and in larger cities 20 %. Then again, it is not clear whether the results of Horanont et al. (2018) can be used in the wide geographical context, as their study is based within one city. The gender ratio or age classes are not used in Telia's extrapolation algorithm.

The regional customer shares are calculated with the following formula:

$$\frac{S_t - S_s}{P}$$

where S_t = Total subscriptions in an area
S_s = Subscriptions in the area that are staying still
P = Population in the area

The reason why subscriptions that are staying still are filtered out from the algorithm is that some people tend to have more than one mobile network subscription. This can mean for example another mobile phone, a television or a tablet, which stay in home all day. By filtering out the subscriptions that are not moving, the total amount of subscriptions can be decreased more near to the actual amount of people that have Telia's subscription. However, this creates an issue regarding personal movement type called immobility. From the results of the NTS it can be seen that not all people leave their apartments during a day. In addition, some people might carry multiple mobile phones with them. These are factors that create variance into the results.

The total amount of subscriptions in an area (S_t) can be calculated either from the customer share information from Telia's customer database (how many subscriptions there are registered in the area) or by observing the amount of connections established in an area each morning (how many subscriptions actually "woke up" in an area and made their first mobile network connection in it). The latter, "first signal based" subscription amount might give more realistic picture of the regional customer share, as the information in the customer database may be faulty. For example, subscriptions owned by other family members may be registered into areas where the actual user does not necessarily live. The study by Bonnel et al. (2018), (see section 2.9), used the same kind of an extrapolation algorithm, as they did not have an access to the mobile network operator's customer database.

After the regional customer share information is generated, the trips and activities are extrapolated with it. A trip is extrapolated with the mean customer share of the origin zone and the destination zone with the following formula:

$$\frac{M_O + M_D}{2}$$

where: M_O = Customer share of the origin zone
M_D = Customer share of the destination zone

The trips between the zones are extrapolated with the resulting average customer share of the origin and the destination by dividing it with the average customer share:

$$\frac{T}{\left(\frac{M_O + M_D}{2}\right)}$$

where T = Number of trips between an origin and a destination
M_O = Customer share of the origin zone
M_D = Customer share of the destination zone

For example, if there are 5 trips between an origin zone which has a customer share of 0,2 and a destination zone which has a customer share of 0,3, the trips are extrapolated as follows:

$$\frac{5}{\left(\frac{0{,}2 + 0{,}3}{2}\right)} = 20$$

In the analyses performed in the thesis work, the customer share information from Telia's customer database is applied instead of the first signal-based customer share.

3.1.5 Aggregation and Anonymization

Aggregation of trips is an obligatory process for conducting an OD-matrix and for ensuring the privacy of people. When studying long-distance transportation, movements of individuals are hard to handle compared to aggregated summarizations. Therefore, individual movement chains are summed up with similar movement chains:

User ID	Origin	Destination	Travel duration
98235	Helsinki	Hyvinkää	52 min
20583	Helsinki	Hyvinkää	39 min
09572	Helsinki	Hyvinkää	44 min
63068	Helsinki	Hyvinkää	45 min
29573	Helsinki	Hyvinkää	58 min

Origin	Destination	Trips	Duration
Helsinki	Hyvinkää	5	0 – 1 h

Figure 7. Demonstration of the trip aggregation process

Aggregation is done to trips and activities both spatially and temporally. The default spatial aggregation levels that are used in Telia's analyses are different sized grids, postal areas and municipalities. These are also the default zones for OD-matrices. However, the default zones can also be joined together for larger areas. Temporally, the time frame of an OD-analysis is set in days, the temporal accuracy of trips and activities being one hour at sharpest. The travel durations can be then again tracked and aggregated also by minutes. Study by Alexandet et al., (2015) proposes that OD-matrices generated from mobile network data are more coherent with national travel surveys when larger spatial aggregation levels are used, but this was not confirmed in the thesis work as small spatial aggregation levels were not used nor validated in the analyses.

There are two methods applied in the Telia's anonymization algorithm (For anonymization, see section 2.10), first of them being ID-shuffle. ID-shuffle means that every user ID is shuffled in the database in every 24 hours. As the possibility of recognizing an individual from his/her travel behavior becomes significantly easier when longer timeframes are accessed (identification of person's home location, work location, and leisure locations is much easier if several days or weeks can be analyzed (Beresfold & Stajano, 2003; Widhalm et al., 2015) Telia uses shuffle interval of 24 hours to ensure that not enough history is available for the identification. This means that if hypothetically one would be able to recognize a person from movement chains of one day, this information could not be used anymore for studying the same person's movements on a next day as the ID:s are shuffled.

At the same time the ID-shuffle makes it basically impossible for Telia to derive home or work locations of users even on an aggregated level.

In addition to ID-shuffle, the k-5 anonymization is applied in the algorithm. This means that no trips are reported unless at least five identical trips are found. For example, if a trip is performed from a cottage to home but no other similar trips are found from the timeframe used in the analysis, the trip count between the zones is marked as zero. However, if the timeframe of the analysis is extended and this way over 4 identical trips are found, the trip count between the locations is marked accordingly. This makes it essential to always consider the aggregation levels used in the analyses together with the k-5 values in order to avoid "k-5 losses". For example, if a postal area level OD-matrix is generated from a single day (For OD-matrices, see section 2.8) and one of the OD-pairs results as 0 due the k-5, the aggregation level can be modified, from a day to a week for instance, and the possibility of capturing more trips for the k-5 lost OD-pair becomes this way significantly higher. The same can be applied also for the zonal aggregations; the postal zone can be widened to the municipality level to eliminate k-5 losses.

3.1.6 Trip Duration Feature

In addition to temporal aggregation of trips into hours and days, trips can be aggregated based on their travel durations. Basically, the most accurate temporal aggregation of trips would this way be gathering the date of the trip, starting time of the trip (in hours), ending time of the trip (in hours) and duration of the trip (in minutes). Same anonymization logic is applied into the travel durations than to everything else; there must be over 5 identical trips to acquire results. Hence, travel durations must be usually aggregated into "travel duration bins", meaning for example 15-minute classes, in order to prevent k-5 losses.

3.1.7 Via-feature

In addition to spatial aggregation of origin zones and destination zones, the trips of OD-matrices generated with mobile network data can be aggregated also by zones passed by along the trip; so-called "via-points". Aim of the ODM-Via analysis is to detect different paths that are used when traveling from origin zone to destination zone and use this information to model route choices and modes of transport. The via-analysis that is used in the thesis work is based on Open Street Map (OSM) nodes. Like described in the section 3.13., Map 5., Telia's trip generation algorithms assign trips into the OSM road network based on the signaling event BSE:s along the way. When the user of the Telia's data processing system wants to set up a via-point, he/she chooses a single OSM street node from the road that is wanted to be used. Then, the algorithm shows the number of trips between an origin and destination that went through that specific OSM link that was chosen.

The Via-point definitions can also be made just by observing the CI-LAC ID:s from the raw signaling event data after specifying which CI-LAC ID:s are along a specific route. If a person is connected to a base station which is located only in the proximity of the specific route that was observed, it is very probable that the person used that route. However, to be sure that all traffic from that route is being observed, one needs to take into account that not all people necessarily connect to every base station along the way. The advantage of the

OSM-based via-analysis is that after a trip is assigned into a specific node, no other base station connections are needed regardless of the location of the actual via-node chosen.

3.1.8 Data Model

The end product of the Telia's mobile network data used in the analyses is a table of trip counts between OD-pairs during a given timeframe. Unlike with National Travel Survey trip table, Telia's data process pipeline produces basically ready OD-matrices, as the aggregations and extrapolations are made already by the data processing algorithms:

Table 5. The data model of Telia's OD-matrix analysis result

Origin	Destination	Trip count
Kuopio	Turku	1233
Seinäjoki	Kuusamo	178
Lappeenranta	Kouvola	743

Regarding the temporal aggregation level used, also the date and the time of trip are included in the data model. If trip duration feature or Via-feature are included in the OD-analysis, these results are generated as an additional column to the right side of the default data model (Table 5.). In an example case of three hypothetical Via-points between Kuopio and Turku, the 1233 trips from Kuopio to Turku are divided into three separate rows, each of them presenting the trip count of one via-point.

3.1.9 Limitations

Despite the relatively large market share that Telia possesses in Finland, the extrapolation based on the amount of mobile network subscriptions without other kind of user background information turns out to be quite difficult. In the general extrapolation methodology, the aim is to acquire samples from as many different demographic groups as possible to be able to widely represent the whole population. However, there is currently no easy way for mobile network operators to do this. It is not only that different demographic groups with different genders and age classes would have different travel behaviors, but they can also have different phone usage behaviors, which can affect the results equally. For example, young people that do not necessarily travel long distances that much might be over-presented in the signaling event logs due the high phone usage, compared to old people who do not use phones that much but might travel more.

3.2 National Travel Survey

The National Travel Survey was used in the thesis work as the main reference data source in validating Telia's mobile network data. The reason for this was that in addition to its comprehensive and relatively large sample size, it is mainly the only data source available

in Finland regarding long trips, modal split (especially regarding rail transport) and travel durations.

3.2.1 Data Gathering

The National Travel Survey gathers basic information about movements of Finnish people (For fundamentals of travel surveys, see section 2.11). This information is being used in developing transport services and infrastructure in addition to promoting traffic safety and decreasing emissions. The survey provides information for the needs of transportation planning regarding modal split and travel purposes. The National Travel Survey is repeated in about six-year intervals to enable up to date data and has been conducted since 1974. Each six-year cycle costs roughly 1 million euros (Finnish Transport Agency, 2013) and is paid with public money. The work is ordered by The Finnish Transport Agency and in 2016 it was delivered by WSP.

The National Travel Survey 2016 timeframe included every day from the year 2016 (1.1. – 31.12.) and additional days from the year 2017 (6.3. – 2.7.) for balancing the answer amounts between seasons (Finnish Transport Agency, 2018b). National holidays, vacation times and weekends were included in the survey. Hence, the survey results represent the movements of the whole year.

Each answerer reported his/her trips regarding one day (00:00 – 23:59), which was assigned to the answerer by the survey management. Additionally, the answerer reported the trips that were over 100 km long and the travels abroad that took place during a three-week period before the actual study date. So for example, if a person was to report every trip that occurred 16.4.2016 (the actual study day), he/she was to report long trips and border crossing trips that occurred between 25.3. and 15.4.2016 (the three-week period before the actual study day).

The National Travel Survey 2016 included national survey section regarding whole Finland and 10 local survey sections regarding largest urban areas in Finland regionally. The national section had a sample size of 22 635 people and the local survey sections had sample sizes between 1000 and 7000 people. The total sample size was 70 215 people. The answering rate of the whole survey was 45 % which resulted as 9307 answerers in the national section and 500 – 3500 in each of the local sections (Finnish Transport Agency, 2018b). In total, there were 31 211 answers in the survey.

The survey section regarding over 100 km long trips was asked only in the national survey section, which resulted as 3314 people reporting over 100 km long trips within Finland. In other words, about 36 % of the total 9307 answerers reported having performed 100 km or longer trips during the given three-week period. 64 % of the answerers did not perform long trips during the study timeframe. This means that every day about 9 people reported their long trips from the preceded three weeks. In total these 3314 people reported 4943 individual long trips in addition to which they were asked to mark down if they repeated the trip during the three-week period, and in the situation of multiple repetitions, how many repetitions there were. With the repetitions summed up with the actual reported individual trips, the National Travel Survey 2016 included 10820 over 100 km long trips within Finland in total, making it 29,6 trips per day in average.

In addition to trip counts, also trip attributes were gathered. The survey questions for short and long trips were not identical. The long-trip survey section questions included origins and destinations, dates, the main mode, trip length (km estimation), trip duration (accuracy of 30 minutes), trip purpose, nights spent at the destination, amount of travel companions and possible optional travel mode.

3.2.2 Representativeness

The National Travel Survey is widened to represent the whole Finnish population. In 2016 this was done with post-stratification, which is a common technique in survey analyses for incorporating population distributions of variables into survey estimates (Little, 1993). The scale of the extrapolation with the National Travel Survey is much more expansive than with Telia's mobile network data, but the survey has significantly more background information of the answerer than mobile network operators. This makes it possible to scale up a single answer in hundreds and still acquire reliable estimates, at least with highly aggregated analyses.

The National Travel Survey 2016 was extrapolated by four variables, them being home location, age class, gender and education. In addition, the study date of the answerer was considered in the extrapolation for balancing the observations for the whole year. The educational background was used in the extrapolation for the first time, as the different answering methods were not divided equally between the educational backgrounds of the answerers. The dataset was extrapolated to represent all the Finnish people that are over six years old, which makes the total amount of 5,1 million people (Finnish Transport Agency, 2018b).

The national survey section was extrapolated separately from the local survey sections, so that the local answers would not have affected the results of the national section. This way all the 22 635 people in the national section were carefully selected regarding their home location, age, gender and education for the widest possible range of attributes. The objective was to have all kinds of people from all around Finland to participate in the study, so that no clear biases would form. After all the people were chosen, each was assigned with an expansion factor, which demonstrated the amount of people that his/her answers would represent.

Like said already above, the survey section regarding over 100 km long trips within the national survey section included 3314 different answerers. The expansion factors of these 3314 people ranged between 96 and 4639, the median being 477 and the arithmetic mean 526. So in average, one trip was extrapolated roughly to 500 trips, depending whether the answerer represented a group with large representation or small representation. With this kind of extrapolation methodology, the 10 820 trips that were observed in the survey were multiplied into about 95,56 million trips in a year. This procedure did not create any kind of new or divergent trips in addition to the ones that were reported but multiplied the reported ones into remarkably larger counts.

3.2.3 Data Model

The end product of the National Travel Survey used in the analyses is a table of long trips. Each row presents one trip and columns present the trip attributes, such as an origin and a destination, the trip length and so on. The unextrapolated trip amount on a row presents the number of trips that the person performed during the three-week study period, and the extrapolated trip amount on the row presents the trips of the all people that the person participated in the study represents. The extrapolated trip amount for each trip in the table is calculated with the following formula:

$$\frac{E * 366 * T}{21}$$

where E = Personal expansion factor of the answerer
T = Total number of trips (how many times the trip was repeated during the three-week period)
366 = Days during the year 2016 (leap year)
21 = Three-week study period length in days

As the National Travel Survey represents the movements of the whole year, temporal aggregation is done by only dividing the total amount of trips into a period wanted. For example, if aiming for a trip amount during a month (31 days), the following formula is used:

$$\frac{T_E}{366} * 31$$

where T_E = Extrapolated trip count

Simplified demonstration of the data model of long trips in the National Travel Survey 2016:

Table 6. The data model of the table of long trips in National Travel Survey 2016

Answerer ID	Origin municipality	Destination municipality	Origin region	Destination region	Repeats	Expansion factor	Trips (year)	Trips (day)
987325	Helsinki	Turku	Uusimaa	Varsinais-Suomi	3	513	35 763	97,7
23758	Lappeenranta	Oulu	Etelä-Karjala	Pohjois-Pohjanmaa	0	1328	23 145	63,2

3.2.4 Limitations

The most significant limitation regarding Finnish National Travel Survey is its sample size. As will be demonstrated in the Analysis parameters -chapter (4.2) and in the Results -chapter, the NTS dataset regarding long-distance trips is barely large enough for any kind of

OD-analyses. Due the sample sizes, directions cannot be profiled between OD-pairs, as this would split the number of observations used for the extrapolation. Same applies to the seasonal changes in travel, which cannot be profiled with the yearly survey. Even though this kind of information can be produced with mobile network data, the validation of those results turns out to be surprisingly difficult as National Travel Survey cannot be used for it.

Another limitation caused by the National Travel Survey regarding OD-analyses was that neighboring spatial zones could not be studied. This was because the NTS section of long trips included only over 100 km long trips, which obviously crossed short regional border crossing trips out. This was considered in the resulting comparison matrices by ignoring the neighboring zones completely (for results, see Chapter 5).

A third limitation which can be mentioned is that NTS does not take in account the yearly traffic amount increasements (The Finnish Transport Agency, 2019) or the portion of non-Finnish people in the traffic. In some road segments of Finland, these variables might play surprisingly significant role.

3.3 Other Reference Data Sources

In addition to the National Travel Survey and the mobile network data itself, data sources that were used in the thesis work included LAM-data, HELMET-model results and Finavia's flight statistics.

LAM (Liikenteen automaattinen mittausjärjestelmä) stands for automated measuring system of road traffic in Finland. It is a network of roughly 500-units of measuring devices all over Finland which aim to gather data on Finnish rubber-tire road traffic (Väylä, 2019). Functioning of a single LAM-point is based on two underground induction loops located consecutively under a single lane. This way it is possible to detect for example vehicle speed, direction, used lane and the length of the vehicle from the temporal difference of the detections between the two loops.

HELMET model (Helsingin seudun työssäkäyntialueen liikenne-ennustejärjestelmän kysyntämalli (Traffic demand system forecasting model of the Greater Helsinki working area)) is a four-step transport model operated by HSL (Helsingin Seudun Liikenne) and delivered by WSP Oy and Strafica Finland Oy (Currently Ramboll). The functions of the model are not described in this thesis work, but the detailed information of the model can be found from the model description (HELMET, 2016). The advantage of using a four-step model in the trip count validation compared to the raw extrapolated trip count information from the NTS is that there are a vast set of parameters in the model that apply the raw survey data into the real world more realistic than the raw data without modeling (HELMET, 2016). For example, information like car ownership, income level, occupation, locations of work places and leisure locations, road capacities and modal split algorithms are added into the equation of trip assignment between locations. In addition, also other datasets like traffic census are used in the modelling besides surveys.

Finavia is a public company in Finland maintaining and developing Finland's airport network. Finavia keeps track on all of the flights operated in Finland and shares some of this data publicly. Table of domestic passenger counts on different airports in Finland (Finavia, 2019) was used shortly in the thesis work for trying to estimate air traffic passenger counts between municipalities.

4 Methods

Whereas the Data & Data Processing -chapter focused on the data sources that were used in the thesis work, Methods-chapter focuses on what is done to them and how. This means explicitly the things that would not have been done without the thesis work. In the project flow chart below, these elements are included in the phases 2 and 3.

First, a project flow chart and clarification of different phases and tools are presented (Figure 8). Then, each of the different OD-matrix analyses is went through by providing the parameters of the analyses and describing the ways how the OD-matrices from different data sources are compared.

4.1 Project flow Chart

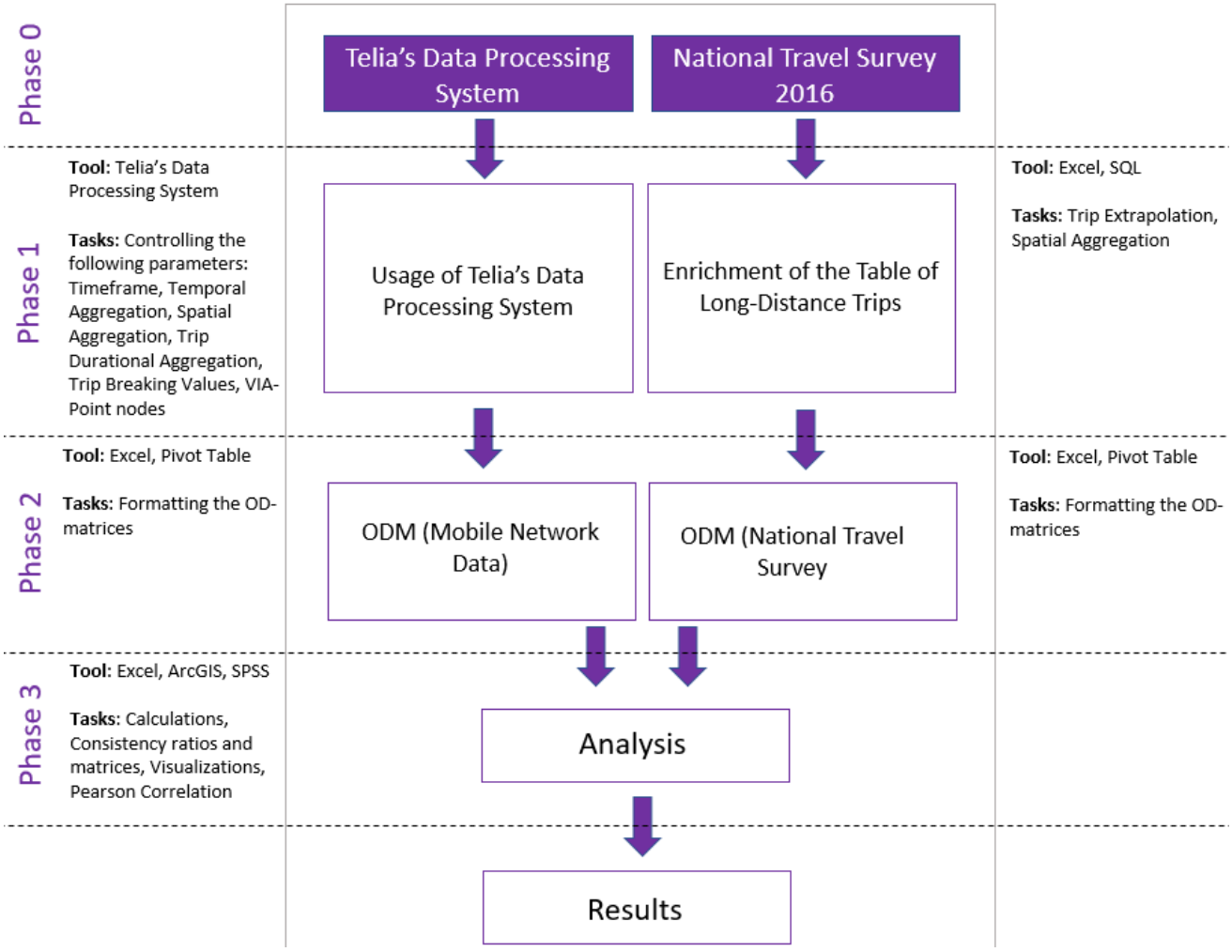


Figure 8. The project flow chart

Phase 0: The starting point of the study. This phase included orientation to the Telia's data processing system and acquiring the National Travel Survey 2016 from WSP Oy, in addition to getting familiar with the related studies of the field. The contents of this phase are mostly

presented in the Data & Data Processing -chapter; how does the Telia's data processing system function, how the NTS is gathered and how it should be used and extrapolated.

Phase 1: Defining the needed OD-matrices and performing the OD-analyses with both datasets. Regarding the mobile network data, this meant giving the input parameters (for example timeframe, spatial aggregation level, temporal aggregation level) to the data processing system and running the analyses with Telia's algorithms in servers specified for that purpose.

Regarding the National Travel Survey, this meant aggregating the reported trips from municipal level to regional and great regional spatial aggregation levels, summing up the trip repetitions and joining the unique answerer extrapolation factors with the reported trips. These tasks were mostly performed with SQL.

Phase 2: Formatting the OD-matrices. The resulting OD-analyses from the phase 1 were formatted with Excel and Excel Pivot Table in the phase 2. Practically this meant simply reformatting a column-based data model into a matrix-based data model. Additionally, as some of the analyses were repeated consecutively for multiple months or multiple days, these matrices were handled and stored accordingly to enable effective analysis in the phase 3.

Phase 3: Validation of the mobile network data. Now when the OD-matrices were generated and formatted in the phases 1 and 2, they were compared and analyzed in different ways. Consistency between the results were studied in Excel with consistency ratios and consistency matrices. The Pearson correlation coefficient (Benesty et al., 2009) was calculated for the datasets with SPSS. Visualizations of different plots and maps were generated with Excel and ArcGIS to present the findings more clearly.

4.2 Analysis parameters

This section contains element from the phases 1 and 3 of the project flow chart (Figure.8). As the ability to compare the results of the mobile network data and the National Travel Survey 2016 are dependent of the parameters that are given for the OD-matrix analyses in the first place, it is necessary to present the reasons behind the parameter choices in the phase 1 to demonstrate the comparison methods in the phase 3. Hence, the aim of the section is to provide information on the parameters of the OD-matrices and on the techniques that were used in the OD-matrix comparison.

4.2.1 OD-matrix parameters

The analyses performed in the thesis work included six different OD-matrices. The main attributes of the six different matrices are given in the table below:

Table 7. Parameters of the OD-matrices generated for the data validation.

Zonal aggregation	Timeframe	Temporal aggregation
Great regions	1.8.2018 – 31.7.2019	Monthly
Regions	January 2019 + July 2019	2 months
Municipalities	1.1.2019 – 31.1.2019	Daily
Municipalities	General weekday	16 weekdays
Municipalities	1.4.2019 – 30.4.2019	Month

Conducting six different OD-matrices meant practically, that each of the mobile network data matrices were generated separately with Telia's data processing system and formatted afterwards with Excel Pivot Table. Parameters regarding extrapolation and breaking values were constant throughout the project; customer data base -based extrapolation and distance dependent breaking value. The parameters regarding zonal aggregation, timeframe and temporal aggregation were given as showed in Table 7.

The National Travel Survey matrix was then again basically the same throughout the project, as the temporal aggregation was done by only dividing the trip amount of the whole year into period desired. Regarding spatial aggregation, the table of long trips of the National Travel Survey was handled just by summing up trips between municipalities to generate trip counts between regions.

4.2.2 Great regional ODM

The most comprehensive and widest OD-matrices were created spatially between "great regions" to create an extensive view into the overall trip count differences in Finland between the datasets. "Great regions" mentioned here and used in the thesis work are not equal to any kind of existing or official administrative areas or areas that would have been mentioned or used in some other context. These areas were created only for the purposes of this study by joining Finnish regions into bigger entities. When talking about great regions in the context of this study, exclusively these areas are meant. These areas should not be confused to other large spatial entities used outside of this study, like Länsi-Suomi etc. or EU: s NUT-areas. The division of great regions used in the study is as follows:

Region		Great region
Uusimaa	1	1
Varsinais-Suomi	2	2
Satakunta	4	4
Kanta-Häme	5	4
Pirkanmaa	6	3
Päijät-Häme	7	5
Kymenlaakso	8	5
Etelä-Karjala	9	5
Etelä-Savo	10	5
Pohjois-Savo	11	6
Pohjois-Karjala	12	6
Keski-Suomi	13	6
Etelä-Pohjanmaa	14	7
Pohjanmaa	15	7
Keski-Pohjanmaa	16	7
Pohjois-Pohjanmaa	17	8
Kainuu	18	6
Lappi	19	8

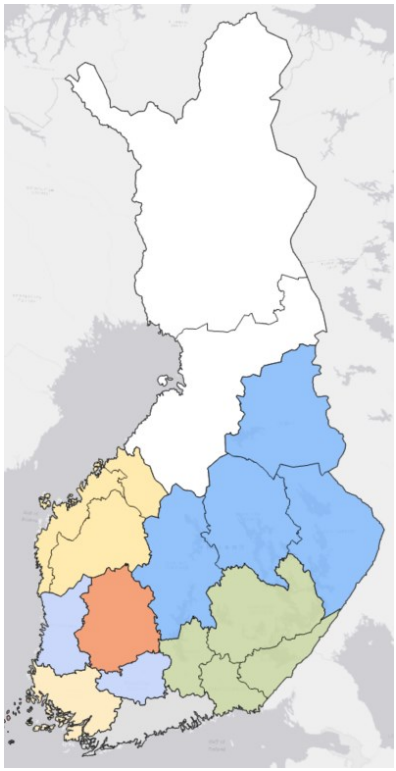


Figure 9. Division of great regions. Great regions are created by joining Finnish administrative regions together to create larger spatial entities. Borders of the administrative regions are shown as black lines on the great regions.

Names of the great regions: Uusimaa (1), Southwest Finland (2), Pirkanmaa (3), Satakunta + Kanta-häme (4), South-Eastern Finland (5), North-Eastern Finland (6), North-Western Finland (7), Lapland (8).

Similarly to the study by Bonnel et al. (2018), the main reason for using these great regions as zones in the OD-matrix is that only with large sized zones the sample sizes of the NTS are large enough for reliable results. This means, that as there are only 10 820 actual reported trips in the survey, they cannot be divided too wide by using small scale spatial aggregation. This would lead into having only couple of observations in each cell of the OD-matrix, which would then again mean that extrapolation and the trip count result of the cell bases only on couple of trips that actually happened. So for example, even though NTS states that there were 50 207 trips between Oulu and Lahti during 2016, the count of trips that actually were reported between Oulu and Lahti was 3. And as we can imagine, 3 trips is a very small number of trips to be reliably extrapolated to represent the movements of every Finnish people during a year between these cities.

Reliability of the NTS OD-matrix can be estimated by reviewing the number of unextrapolated trips per OD-cell:

NTS Great regional ODM, reported trips

	1	2	3	4	5	6	7	8
1								
2	791							
3	604	239						
4	562	251	236					
5	1231	74	189	104				
6	573	71	257	93	403			
7	230	78	199	78	35	196		
8	194	64	85	43	68	383	161	

Figure 10. Number of unextrapolated trips in the NTS great regional ODM, 1.1.2016 – 31.12.2016. Green: Over 100 trips, Yellow: 50 – 100 trips, Red: under 50 trips.

With the great regional spatial division, the reported trips in the National Travel Survey formed relatively good sample sizes. In most of the cells there were over 100 trips (Figure 10.) which seemed to be large enough sample sizes, at least in the context of the thesis work results. The sample sizes between 50 and 100 trips seemed to be mostly reliable, but also an exception was observed. The sample sizes below 50 trips could not then again be considered reliable, even though the other one of the cells (for the results, section 5.1) provided coherent results with the mobile network data.

The great regional ODM by Telia's mobile network data was created by processing each month separately and aggregating all the trips between great regions into a sum of the according month. After all the separate months were generated, the monthly trips were summed up with each other to create the yearly ODM. This way the resulting yearly OD-matrix presented the trips of the whole year within Finland between great regions. The yearly ODM by National Travel Survey was then again formed similarly to the Figure 10., but with the extrapolated trip counts.

The actual comparison between these two great regional OD-matrices was done by calculating a consistency ratio for each of the OD-pairs. The result was this way an OD-matrix, in which there were consistency ratios in each of the OD-cells. The ratio showed whether the trip counts between the two data sources were near each other. The results are described in section 5.1.

4.2.3 Regional ODM

Similarly to great regional ODM, also regional ODM was generated for the validation. Timeframe of the regional ODM was one year and zonal aggregation was done by Finnish administrative regions. For the National Travel Survey this meant that the origins and the destinations were aggregated into regions instead of municipalities. Regarding Telia's mobile network data, one-month ODM analysis between regions was performed for January and also for June. Then these two months were summed up and the results multiplied by six, in order to acquire trip counts that would represent the "average year" as the NTS results did. The trip counts of this kind of synthetic year were compared to the results of the great regional ODM by mobile network data in which the whole year was processed. The

differences between trip counts were so insignificant, that the "(Jan+Feb)*6" was allowed to represent the whole year in this "secondary" comparison.

The reason why the regional ODM comparison is called secondary is that it turned out to be basically unusable. The problem with the regional ODM was that the sample size of National Travel Survey 2016 was not large enough. When 10 820 trips are distributed into 171 OD-cells, the average sample size per cell is 63 reported trips. Compared to the sample sizes per cell in the great regional ODM (Figure 10.), even the 63 trips per cell would do just fine, but as the trips are nowhere near equally distributed between all the cells, most of the cells end up unusable:

NTS regional ODM, reported trips

	1	2	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1																		
2	791																	
4	184	164																
5	378	87	39															
6	604	239	148	88														
7	533	28	10	27	92													
8	329	22	17	20	40	19												
9	148	9	0	15	23	10	97											
10	221	15	1	14	34	59	42	52										
11	138	13	7	8	46	13	16	14	85									
12	113	2	4	10	17	14	13	23	46	99								
13	291	47	27	35	190	33	23	11	78	111	18							
14	125	35	28	19	148	2	3	2	11	16	4	117						
15	65	38	19	2	38	7	2	0	0	4	2	13	88					
16	17	5	9	1	13	2	0	0	6	3	0	12	25	25				
17	137	36	19	6	56	18	11	1	13	74	16	63	55	42	36			
18	31	9	2	0	4	1	5	5	2	49	20	9	20	3	2	150		
19	57	28	7	11	29	5	2	16	2	19	9	22	15	8	5	324	30	

Figure 11. *Number of unextrapolated trips in the NTS regional ODM, 1.1.2016 – 31.12.2016. Green: Over 100 trips, Yellow: 50 – 100 trips, Red: under 50 trips.*

4.2.4 Municipality ODM, daily

The third OD-matrix was generated with a one-month timeframe and a daily temporal aggregation. The spatial aggregation of the third ODM was based on municipalities, which was even smaller zonal scale than with the regional ODM. This obviously meant that the ODM based on municipalities could not be validated comprehensively with the National Travel Survey 2016, but instead the aim was to choose specific OD-pairs with relatively large sample sizes in the NTS and study how accurately Telia's mobile network data is able to profile daily differences in trip counts during a month.

For the specific OD-pair Helsinki metropolitan region (Helsinki, Espoo, Vantaa) and Turku were chosen. With this kind of municipality combination, 259 reported trips between the zones in the NTS were the basis of the validation. After extrapolation, these 259 trips formed 3 044 552 trips in a year, which meant 8 318 trips per day. Unlike with the great regional ODM and the regional ODM, the 8 318 trips were divided by two to acquire trip counts per

direction, so that capabilities of Telia's mobile network data in detecting daily trip count changes between directions could be assessed.

A major problem that raised when particular days were tried to be validated with the NTS yearly average, was that seasonal differences in long distance transportation are so remarkable, that a yearly average trip count is way too high for winter days and way too low for summer days. Regarding the results of the thesis work (see section 5.3), there are 24,7 % less trips between Helsinki and Turku during January than the yearly average states. Hence, in order to decrease the NTS results regarding validation of mobile network analysis on January, the results of the NTS were multiplied by 0,753 to acquire more realistic winter time trip count average.

4.2.5 Municipality ODM, general weekday

The fourth OD-matrix was created for the validation of medium length trips between municipalities of Greater Helsinki. This matrix differed from the previous matrices in couple of ways, the most significant difference being the reference data source. The ODM of Greater Helsinki was not validated directly with the results of the National Travel Survey 2016, but with a transport demand model called HELMET. Even though this validation method added more moving parts in to the validation equation and this way made the recognizing of potential root cause errors more difficult, it added some notable elements into the NTS data and made it this way more realistic.

As HELMET model represents the traffic on an average weekday during fall, this kind of an average weekday had to be produced also with Telia's mobile network data. This was done by aggregating 16 weekdays from the September of 2018 into a sum of one OD-matrix and then dividing the resulting trip counts by 16. Unlike with the other OD-matrices of the thesis work, the ODM of Greater Helsinki enabled including short trips into the analysis and made it this way possible to study trip counts also between neighboring municipalities. This was not possible with the previous OD-matrices, as the NTS:s table of long trips did not include trips shorter than 100 kilometers. The actual comparison between the ODM generated with HELMET and the ODM generated with mobile network data was done similarly to the previous comparisons; by calculating consistency ratios and absolute trip count differences for each of the OD-cells.

The OD-matrix generation with HELMET-model was not done as a part of the thesis work project. The OD-matrix with trip counts by HELMET between Greater Helsinki municipalities was delivered by Strafica, as an already ready-to-use matrix.

4.2.6 VIA + Duration ODM:s

As all of the OD-matrix analyses that have been presented this far have aimed for validating the consistency of trip counts of Telia's mobile network data and other data sources, the OD-analysis called "Via + Duration" moves from the trip counts to the detection of route choice and modal split.

The OD-matrices that were generated for the evaluation of route choice estimation capabilities and modal split detection capabilities were created with a timeframe of April 2019 and a temporal aggregation of one month. For spatial aggregation, municipality level

was used. However, regarding Helsinki area, Helsinki, Vantaa and Espoo were combined in order to acquire more unextrapolated trips from the NTS for more reliable reference.

The original objective of the thesis work regarding modal split was to validate existing modal split algorithms developed already by Telia. These algorithms were developed to consider other spatial data references in the transport mode detection, such as airport locations, railway station locations, rail track links, etc. The aim of the algorithm development was to create and validate an algorithm which was standardized and enriched with the other spatial data and which was this way able to detect a mode of transport by default, without the user having to determine parameters for each OD-pair separately. However, at the time of the thesis work project, this algorithm was still under development and could not been used.

Forced by these circumstances, the thesis work project ended up proposing an alternative rule-based method for detecting a mode of transport. This was done by combining the trip duration feature (see section 3.1.6) and via-feature (see section 3.1.7) to recognize different geographical trip paths and travel speeds. By observing each OD-pair individually, it is usually possible to find differing geographical paths used by trains and rubber-tire vehicles between OD-pairs. Hence, if a rail track goes far away from the highways, the locations of the signaling events in the proximity of rail track or highway are relatively clear evidence of the mode used. Then again, if the distance travelled is long enough, air trips can be detected from the travel durations. For example, it is practically not possible to fly from Helsinki to Oulu in 8 hours, nor it is possible to drive from Helsinki to Oulu in an hour.

With the combination of travel duration feature and via-feature, modal split was detected for two individual OD-cases, the first one being Helsinki – Kuopio and the other one being Helsinki – Lappeenranta. In addition, the via-feature was utilized for detecting a trip count division between different highway route options regarding the two OD-pairs mentioned. The more accurate usage of the features is presented in the Results -chapter, section 5.4.

5 Results

This chapter will demonstrate the results of the thesis work. At the beginning of the chapter the main emphasis will be on validating the results of the mobile network data and indicating the correlation achieved between OD-analyses by mobile network data and the National Travel Survey. However, also additional insights achieved by mobile network data regarding long distance transportation are presented besides the validation, and these insights will be the emphasis in the ending of the chapter.

5.1 Great Regional ODM

The great regional ODM-analysis was the most comprehensive and important analysis in the thesis work. This analysis set the foundation for all the other analysis to come, as it presented the correlation between mobile network data and the National Travel Survey in the big picture. Without achieving high correlation regarding Finnish travel behavior between great regions, the more specific analyses would have been useless. However, coherent results were obtained.

5.1.1 Great Regional ODM, Consistency

Like described in the section 4.2.2, the great regional ODM validation was done by calculating consistency ratios for each of the OD-cells, excluding the cells which represented neighboring great regions. The neighboring cells were disregarded due the inability of the NTS to observe short trips. Each consistency ratio is calculated by dividing an NTS based OD-pair trip count result with the mobile network data -based OD-pair trip count result. The resulting consistency ratios are read as follows:

Consistency ratio < 1 = The mobile network data -based trip count was higher than the NTS-based trip count
Consistency ratio = 1 = The mobile network data -based trip count was equal with the NTS-based trip count
Consistency ratio > 1 = The mobile network data -based trip count was lower than the NTS-based trip count

The total trip counts between unneighboring great regions during a year were about 22,15 million (NTS) and 22,65 million (Telia's mobile network data). The resulting consistency matrix is given below:

Trip amount ratio per OD-pair

	1	2	3	4	5	6	7	8
1								
2	N							
3	0,8600325	N						
4	N	N	N					
5	N	1,06437	N	N				
6	1,2402091	0,94638	N	0,86519	N			
7	0,8167577	0,92454	N	N	0,70198	N		
8	0,8742808	2,196	1,01465	1,10726	1,085	N	N	

Total Trip Amount Ratio

1,0227747

Figure 12. The result matrix of trip amount ratios between great regional OD-pairs. Green: consistency ratio 0,8 – 1,25, Yellow: consistency ratio 0,7 – 0,8 or 1,25 – 1,43, Red: consistency ratio below 0,7 or above 1,43. "N" stands for "Neighbor" which means neighboring great regions.

Consistency ratio classes (green, yellow, red) were defined in a way that class boundaries were the counter values of each other. This way the relational differences into each direction (too many trips compared to the NTS or too little trips compared to the NTS) were equal. As seen in the Figure 12., 11/13 of the OD-pairs achieved a consistency ratio between 0,8 and 1,25, which is in the context of the thesis work considered as a match. Correlation of 0,975 was obtained between the 13 cells in total (Table 8.) The total trip amount ratio (22,65 million / 22,25 million) is shown in the lower left corner (1,02).

Table 8. Pearson correlation between the great regional ODM by the National Travel Survey 2016 and Telia's mobile network data.

Correlations

		NTS	Telia
NTS	Pearson Correlation	1	,975**
	Sig. (2-tailed)		,000
	N	13	13
Telia	Pearson Correlation	,975**	1
	Sig. (2-tailed)	,000	
	N	13	13

**. Correlation is significant at the 0.01 level (2-tailed).

Even though the overall consistency between the datasets mostly matches, there are two OD-pairs with inconsistent results. Regarding OD-pair 5 – 7 (South-Eastern Finland – North-Western Finland) there are 43 % more trips generated by the mobile network data compared to the NTS. Then again with the OD-pair 2 – 8 (Southwest Finland – Lapland), the mobile network data gives only half of the trips compared to the NTS. The most likely reason for this inconsistency seems to be too small sample sizes in the unextrapolated trips of the NTS. These two inconsistent OD-cells are the lowest and third lowest (Figure 13.) unextrapolated trip count cells between great regional OD-pairs in the NTS:

Total observed trips in National Travel Survey 2016

	1	2	3	4	5	6	7	8
1								
2	791							
3	604	239						
4	562	251	236					
5	1231	74	189	104				
6	573	71	257	93	403			
7	230	78	199	78	35	196		
8	194	64	85	43	68	383	161	

Figure 13. *Number of unextrapolated trips in the NTS great regional ODM, highlighting the cells (red) where the resulting trip counts did not match with Telia's mobile network data.*

There are no clear outliers in the long trip table of NTS regarding these particular inconsistent cells that would influence the result remarkably. An outlier could be for an instance a person with high personal expansion factor who would have travelled between Turku and Lapland irregularly often during the three-week study period. However, the observations regarding these inconsistent OD-pairs were in line with all the other OD-pairs in the table of long trips. Still, as 85 % of the cells provided coherent results and the disparate results landed only on the cells with poor sample sizes, it is likely that the errors are caused by faulty trip counts of the NTS. Same observation was acquired regarding the Regional ODM; the probability for inconsistency between data sources was proportional to the scarcity of unextrapolated trips in the NTS. By combining great regions 7 (North-West Finland) and 8 (Lapland) and reviewing the trips again between 7+8 and 2 (Southwest Finland), the consistency ratio decreases from 2,2 to 1,247 and classifies this way as a match. This observation also promotes the assumption of the error caused by too small sample size of the NTS, as the result can be improved by increasing the sample size.

5.1.2 Great Regional ODM, Total Seasonal Changes

As the great regional ODM was ran for a whole year, this enabled presenting the seasonal changes in travel with the mobile network data. The analysis was run with a monthly temporal aggregation, so trips during each month were summed up. For validation, the NTS:s long trip table was used by dividing the yearly trip count by 12 and having the resulting trip count (1,85 million) as a constant reference for every month. Similarly to the consistency validation (see section 5.1.1), only unneighboring great regions were used:

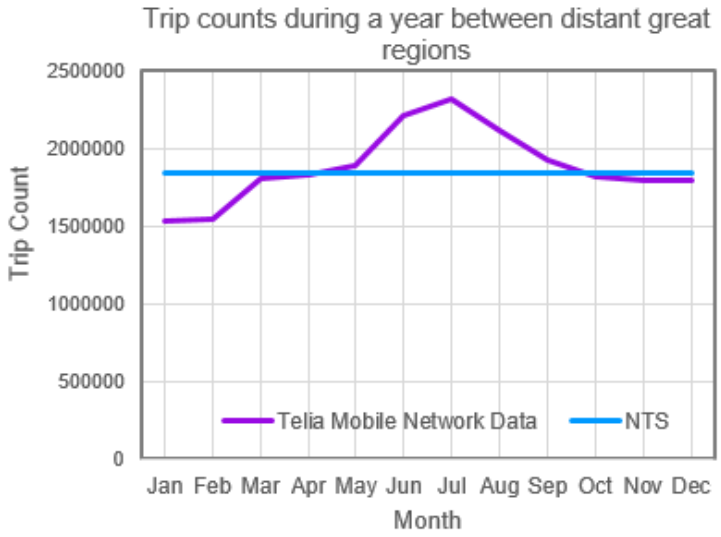


Figure 14. Seasonal changes in long distance travelling by Telia's mobile network data. The sum of trip counts between unneighboring great regions is calculated for each month.

Telia's mobile network data provides promising results regarding seasonal changes (Figure 14.). The results seem to be in line with the general assumption of summers standing out with higher trips counts compared to other seasons and fall standing out above January and February. There are no valid reference data sources in Finland regarding trip count changes between seasons, but road traffic flow counts can be used to estimate overall travelling during a year. A LAM point (Liikenteen automaattinen mittausjärjestelmä - automatic traffic measuring device) 1601 (vt1 Lakiamäki) was chosen to be compared to Telia's great regional ODM results. Telia's data presents trips between Uusimaa and Varsinais-Suomi and the LAM 1601 is located in the middle of those areas at Helsinki-Turku highway. Similarities between data sources can be seen below (Figure 15.):

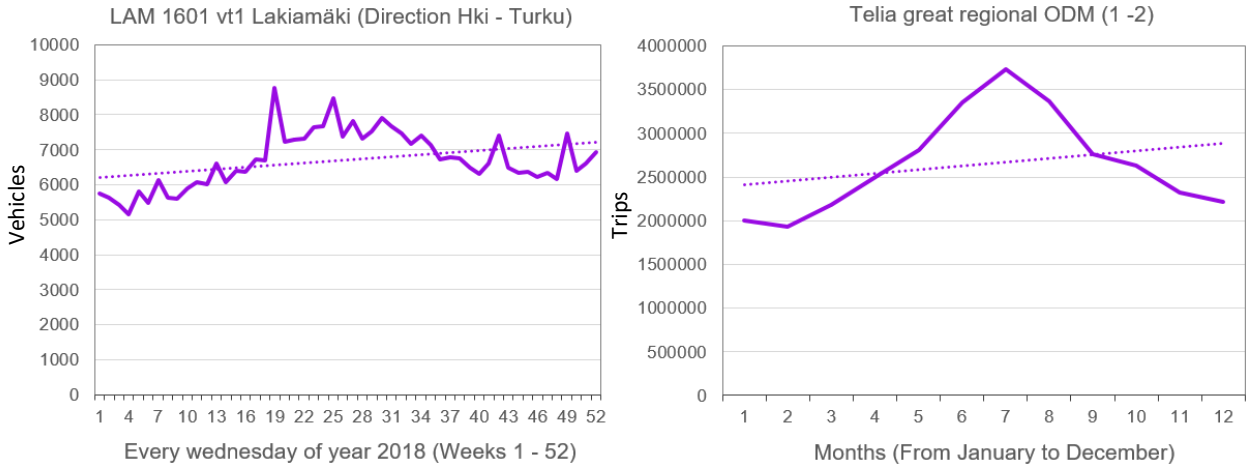


Figure 15. Demonstration of Finnish seasonal travel patterns: low level winter, high level summer and medium level fall. Road traffic count on LAM 1601 vt1 Lakiamäki on every Wednesday during 2018 (left) and Telia's OD-trip count between Uusimaa and Varsinais-Suomi from January to December (right).

Like the LAM point presented above (Figure 15.) also other LAM points all over Finland confirm the three clear phases in Finnish long distance travelling: low level winter, high

level summer and medium level fall. Hence, Telia's great regional ODM (Figure 14.) seems to provide a reliable picture of the Finnish travel behavior between seasons.

5.1.3 Great Regional ODM, Differences Between Seasonal Changes

Even as the total trip count profile between distant great regions (Figure 14.) gives an overall picture of the seasonal changes in Finnish travelling, there are remarkable differences between individual OD-pair trip count profiles when it comes to the seasonal changes. For fully being able to understand the traffic amounts on certain links during a given timeframe, these OD-pair trip count profiles must be reviewed individually.

Regarding the thesis work results, seems to be that even though the three clear trip count phases (low level winter, high level summer and medium level fall) can be found from most of the OD-pair profiles, some of the summer high peaks and winter low peaks can differ noticeable from the average, whereas some OD-pairs might show only small variance throughout the year. Some areas tend to provide more activities than others in different times of the year, like Central-Finland during the summer cottage season and Lapland during winter. One reasonably general trip count profile (Southwest Finland (2) – South-Eastern Finland (5)) and one sharp peaked summer profile (Southwest Finland (2) – North-Western Finland (7)) are presented below (Figure 16.):

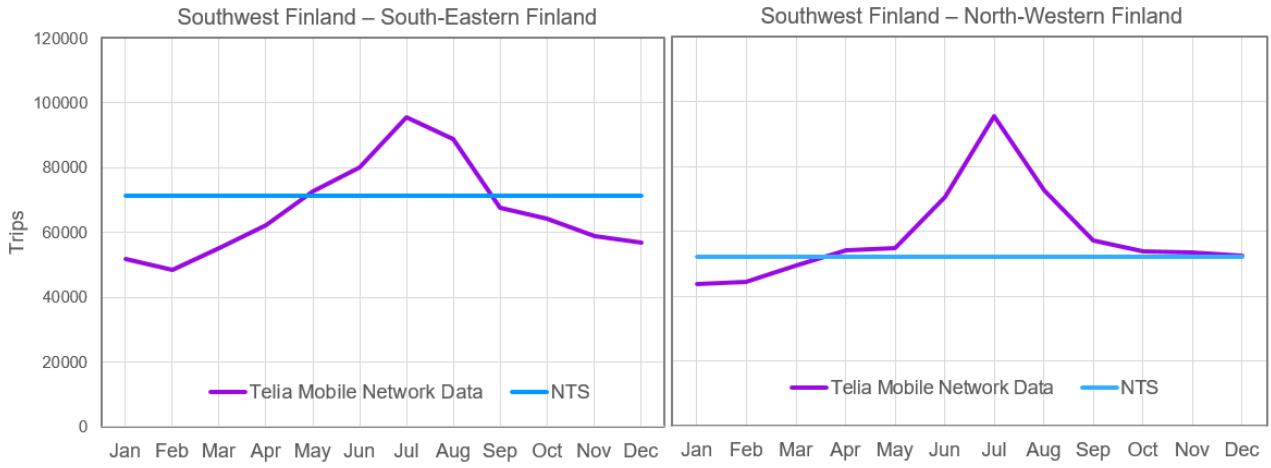


Figure 16. Trip count profiles of Southwest Finland – South-Eastern Finland and Southwest Finland – North-Western Finland. As a reference, the NTS yearly trip count result between the zones divided by 12.

There are two main problems when presenting trip counts between two locations based only on yearly average. Firstly, if a yearly trip count profile is curvy shaped, a straight line representing the average trip count per month does not necessarily represent the reality at all. For example, as seen from the graph in left side of the Figure 16., the trip count between Southwest Finland (2) and South-Eastern Finland (5) starts from the low-level winter, rises up to the summer high peak and then decreases back down below the average. However, even though the total trip count matches between Telia's mobile network data and the NTS, the NTS:s monthly average matches with the curvy trip count line only at May and September. This means that the trip count given by the National Travel Survey is correct only two times during a year.

Secondly, in the context of the thesis work results, seems to be that the OD-pairs with a sharp summer peak trip count profile are extrapolated incorrectly regarding seasonal differences in the National Travel Survey 2016. In addition to the high summer peak trip count profile that is presented in the right side of the Figure 16., also two other similar trip count profiles were found by the mobile network data; Southwest Finland (2) – North-Eastern Finland (6) and Satakunta + Päijjät-Häme (4) – North-Eastern Finland (6). Like seen in the Figure 12., consistency ratios of these three OD-pairs are below 1, meaning that the mobile network data provided higher total trip counts than the NTS. Then again, the OD-pairs which did not have as sharp summer peak did not have this kind of a correlation in the consistency ratios. This observation promotes the assumption that a sharp summer peak (like in the right side of figure 16.) is difficult to capture with a small sample sized survey and this leads into underestimating the total trip count between those locations.

It might also be misleading to compare the yearly trip count averages from the NTS between different OD-pairs without considering the differing effects of seasons into different locations. Even though the total yearly trip count profile (Figure 14.) shows the total profile of Finnish long distance travelling during a year, all the individual OD-pair trip count profiles do not necessarily behave accordingly. Relations between some OD-pairs might change between seasons, like between Uusimaa (1) – North-Western Finland (7) and Uusimaa (1) – Lapland (8):

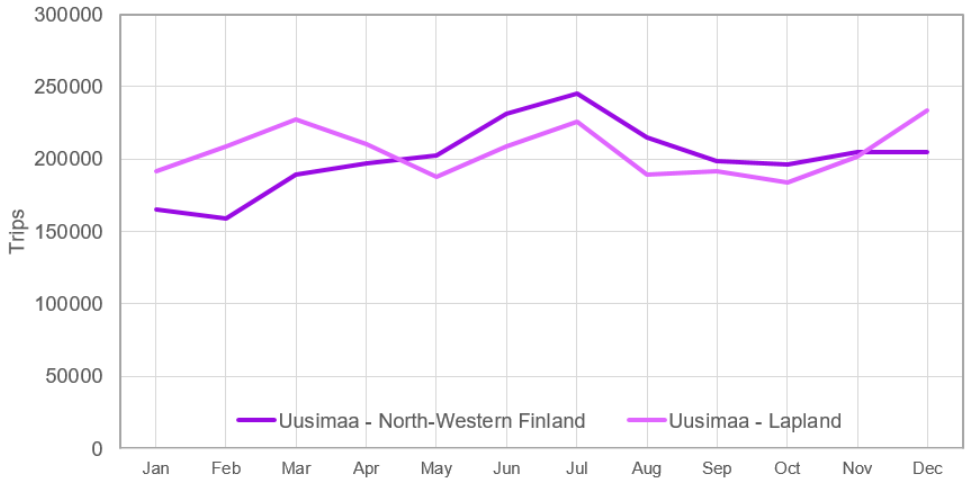


Figure 17. Telia's mobile network data trip counts during a year between Uusimaa (1) – North-Western Finland (7) and Uusimaa (1) – Lapland (8) show that some OD-pair trip count profiles behave differently compared to each other.

Even though there are more trips during a year between Uusimaa (1) – Lapland (8) (NTS: 2147, Telia: 2456) than between Uusimaa (1) – North-Western Finland (6) (NTS: 1947, Telia: 2423), the trip count from Uusimaa to North-Western Finland (6) is higher than the trip count from Uusimaa to Lapland (8) seven months a year, which is more often than the other way around. The main reason for this is that Lapland is a popular holiday destination also during winter (peaks at winter holiday seasons), whereas North-Western Finland has more regular yearly trip count profile, with trips focusing on the summer season.

5.2 Municipality ODM, HELMET

In addition to great regional ODM, also Greater Helsinki ODM results were used in the consistency validation of Telia's mobile network data. Comparison between the results of HELMET four step transport demand model and Telia's mobile network data gave possibility to validate also shorter trips between municipalities. As the analysis was not tied into the limitations of the National Travel Survey, trip counts per directions were also distinguished. Like with the Great Regional ODM and the National Travel Survey, also the results of the comparison between HELMET and Telia's municipality ODM gave promising results.

The consistency between the datasets is presented in the same format than with the great regional ODM, with one exception (for reading a consistency matrix, see section 5.1.1). Regarding all the OD-pairs where Helsinki, Espoo or Vantaa are not included, the trip count differences are presented with absolute trip count differences (Telia – NTS) and not consistency ratios like in the great regional ODM consistency matrix. This is because small municipalities have very low trip counts between each other, and for this reason it is not optimal to compare them with a ratio. For example, HELMET model shows that there are two trips from Kauniainen to Pornainen during an average weekday in September, whereas Telia's mobile network data shows that there is only one. Even though the actual trip count difference between the datasets is only one trip, in the name of consistency ratio there is a double amount of trips by the mobile network data compared to HELMET. This would classify as a very bad consistency ratio even though the difference is only one trip. A black line is drawn into the matrix to distinguish the values that are consistency ratios and the values that are absolute trip count differences:

Trip Amount Ratios per OD-pair and Absolute Trip Count Differences

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1		0,96	0,99	1,01	0,92	1,27	1,51	1,32	1,56	1,38	0,67	0,92	0,97	1,51
2	0,99		0,92	1,59	0,7	0,93	0,87	0,9	1,16	1,09	0,66	0,71	0,89	1,26
3	0,99	0,91		1,22	0,78	0,85	0,77	0,6	1,14	1,2	0,61	0,72	0,74	1,23
4	0,98	1,56	1,16		96	16	-23	-2	-8	-6	9	1	4	-1
5	0,94	0,76	0,79	93		-601	-136	-38	-55	-33	16	-20	-37	-13
6	1,35	1,05	0,93	5	-676		-760	-156	-38	-99	-24	-57	-590	-14
7	1,47	0,87	0,83	-23	-156	-589		-669	-815	-1251	-91	-245	-2607	-53
8	1,22	0,84	0,61	-4	-28	-136	-413		-2761	-826	-419	1850	-1880	-288
9	1,56	1,18	1,1	-5	-57	-32	-725	-2802		-4142	1462	-596	-323	-464
10	1,48	1,12	1,2	-6	-38	-88	-1141	-696	-4495		131	-1649	-1354	-753
11	0,63	0,64	0,63	12	12	-20	-92	-496	882	-44		143	-80	735
12	1,02	0,78	0,76	-1	-29	-54	-243	1785	-779	-2080	124		-1767	337
13	1,1	0,93	0,79	1	-43	-531	-2306	-1941	-404	-1512	-90	-1686		-104
14	1,53	1,35	1,25	-1	-14	-15	-58	-335	-553	-943	651	300	-111	

Total Trip Amount Ratio

1,02

Average of the Absolute Trip Counts

-397

Figure 18. Trip amount ratios per OD-pair and absolute trip count differences between HELMET model and Telia's mobile network data. Trip amount ratios are generated for the OD-pairs which include Helsinki, Espoo or Vantaa, and absolute trip count differences for the rest of the OD-pairs.

Trip amount ratio classes: Green: 0,75 – 1,33, Yellow: 0,6 – 0,75 and 1,33 – 1,66, Red: below 0,6 or above 1,66.

Absolute trip count difference classes: Green: -200 – 200, Yellow: -1000 – (-200) or 200 – 1000, Red: below -1000 or above 1000.

Zones: Helsinki (1), Vantaa (2), Espoo (3), Kauniainen (4), Kirkkonummi (5), Vihti (6), Nurmijärvi (7), Tuusula (8), Kerava (9), Järvenpää (10), Sipoo (11), Mäntsälä (12), Hyvinkää (13), Pornainen (14).

The consistency matrix (Figure 18.) between HELMET and Telia's mobile network data shows that only 10 % of the OD-cells result inconsistent. Then again, 90 % of the cells are either complete matches (green) or roughly in the same range (yellow), the portion of complete matches being 60 % from the total cell count. Remarkably high correlation of 0,996 is obtained between the datasets (Table 9.). Also, as the inconsistent cells are not randomly spread, but tightly bound into seven particular small municipalities, the result seems to promote a high reliability and validity of Telia's mobile network data in the big picture. The total trip amount ratio indicates complete match in the overall picture, even though the average of the absolute trip counts (the arithmetic mean of all the cells with an absolute trip

count difference) shows that regarding the small municipalities, HELMET generated 397 trips more in average per cell than Telia's mobile network data.

Table 9. Pearson correlation between the HELMET municipality level ODM and Telia's mobile network data.

Correlations

		NTS	Telia
NTS	Pearson Correlation	1	,996**
	Sig. (2-tailed)		,000
	N	182	182
Telia	Pearson Correlation	,996**	1
	Sig. (2-tailed)	,000	
	N	182	182

**. Correlation is significant at the 0.01 level (2-tailed).

5.3 Municipality ODM, daily

The second municipality scaled ODM was generated with a timeframe of one month and a temporal aggregation of one day. The aim of the matrix was to profile daily differences in trip counts and detect differences in trip counts regarding directions. However, overall consistency between Telia's mobile network data to other data sources was not measured due the lack of reference data.

The daily trip count profile of January 2019 between Helsinki region (Helsinki, Espoo, Vantaa) and Turku is given below:

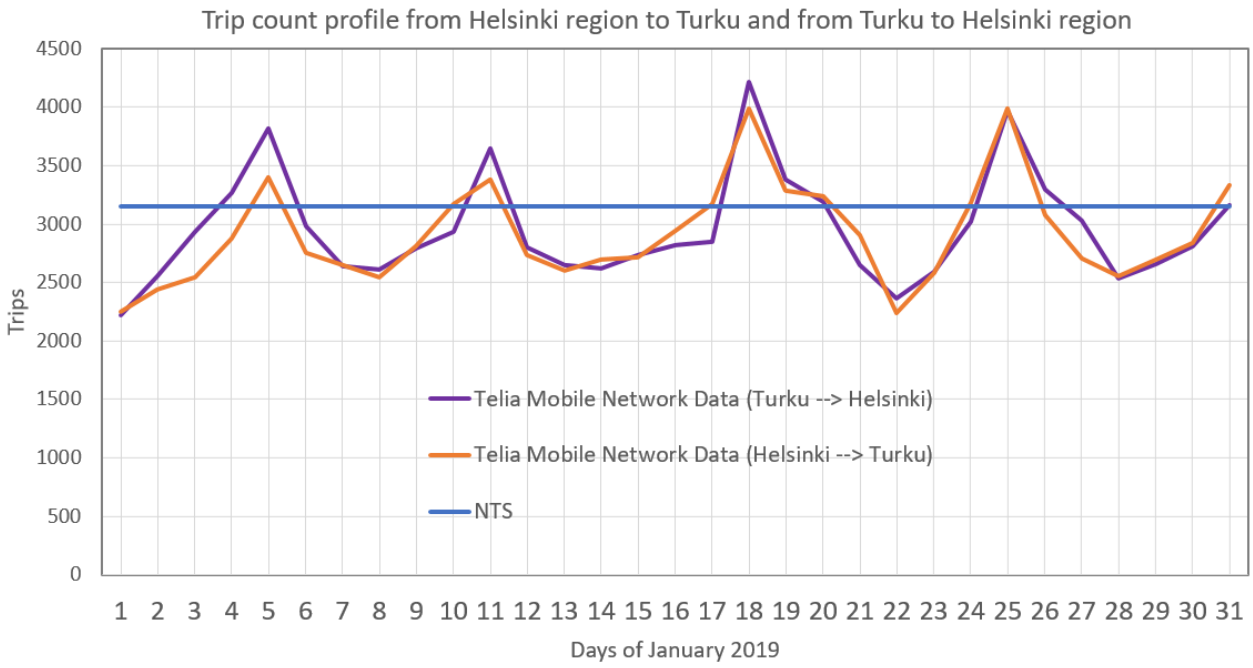


Figure 19. The daily trip count profile of January 2019 between Helsinki region (Helsinki, Espoo, Vantaa) and Turku

As seen in the Figure 19., mobile network data seems to be able to detect changes between daily trip counts and directions quite easily. After winter break and new year holidays people are returning from Turku to Helsinki in larger counts than normal. The typical high peak day is Friday, but the first high peak of the year is Saturday, due the winter holidays. Low peak occurs weekly at Wednesdays, even though the first day of the year (Tuesday) is low also. As Turku profiles as a city with a university and business attractions, no clear differences occur between directions excluding national holidays and special events, like new year. Then again, as the analysis was repeated for Kuopio for instance, significant directional differences were detected between Fridays and Sundays, as people travel to spend weekends in the urban metropolitan region or the other way around spend summer weekends in Kuopio. This kind of differences are easy to confirm with the LAM points which detect directions separately, even though the road traffic counts are not directly comparable to OD-trip counts in total quantities.

5.4 Route Choices, Travel Durations, Modal split

Whereas the first part of the Results -chapter validated the ODM trip counts as the foundation for deeper level analyses, this section dives into the route choices, travel durations and modal split. As the Methods -chapter already demonstrated the fundamentals of these analyzes (see sections 4.2.5 and 4.2.6), this section combines them all in order to generate modal split for individual case OD-pairs.

This section proposes a method to combine via-point information and travel durations for distinguishing road traffic, rail traffic and air traffic. This method is based on two individual OD-pair cases, in which the method was tested. First, Helsinki – Kuopio OD-pair is presented, which ended up showing great consistency with the National Travel Survey including air travel separation. Then we move to Helsinki – Lappeenranta OD-pair, which

shows the capabilities of mobile network data being able to detect rail traffic even better than the National Travel Survey.

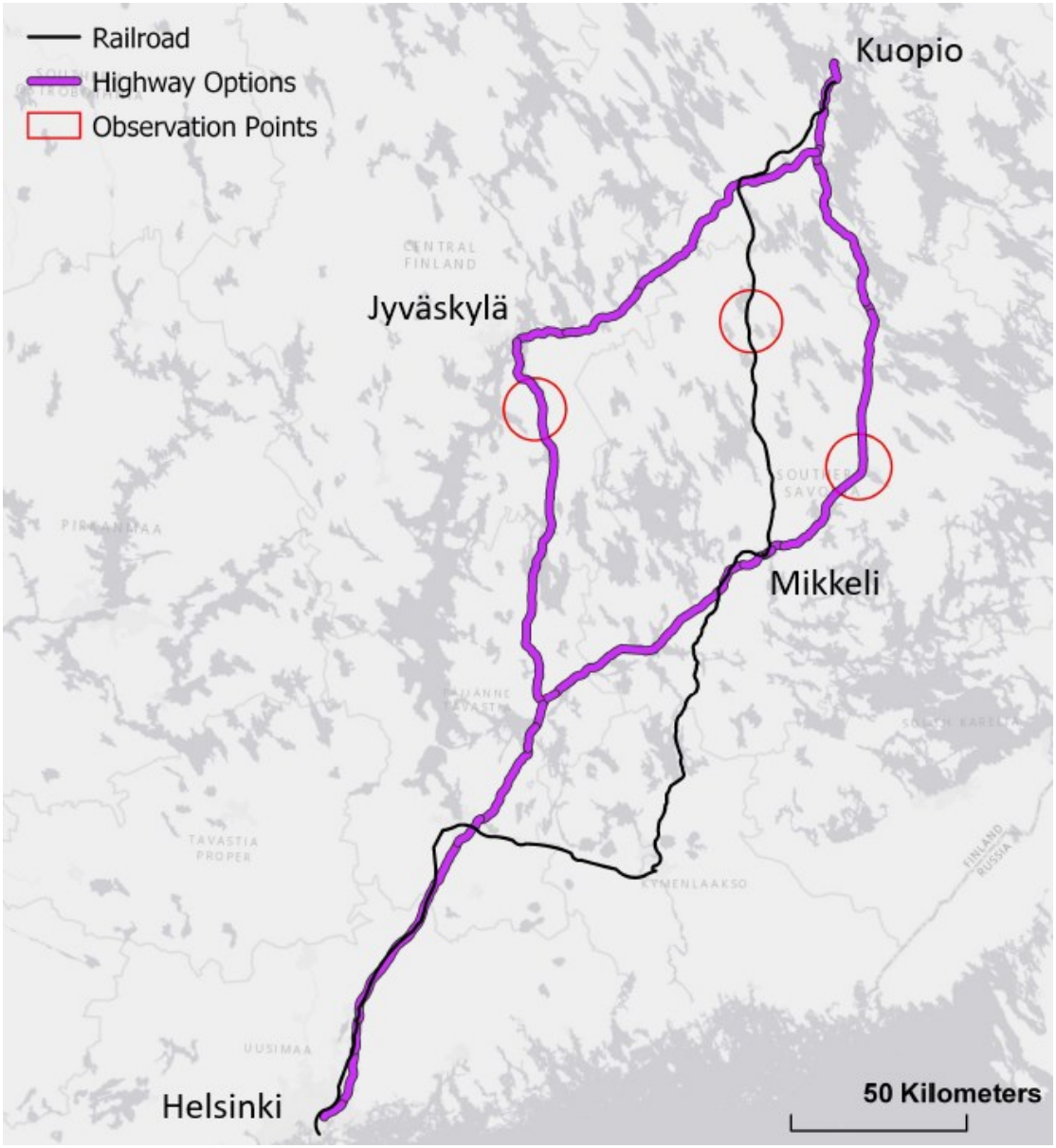
5.4.1 Case Helsinki – Kuopio

The first case OD is Helsinki – Kuopio. To be more precise, Helsinki stands for Helsinki, Espoo and Vantaa, as this larger spatial aggregation enables us to have more NTS observations and this way more reliable validation data reference. Telia's ODM was ran with monthly spatial aggregation, the timeframe of the analysis being 1 – 30 April 2019. The unextrapolated trip count of NTS between these zones was 51 trips made by 30 persons. Like seen from the earlier parts of the thesis work results, this is not a good number of reference trips, but under the circumstances still the best reference available. The limit of 50 trips can be considered to be as a decent amount of unextrapolated reference trips, which is in this case reached. Telia's market shares (without company subscriptions) that are used for the mobile network data extrapolation are roughly 20 % for Helsinki, Vantaa and Espoo and 15 % for Kuopio.

The analysis was started by going through different highway route options from Helsinki to Kuopio based on Google Map navigator results. The most suitable route options by Google were route via Mikkeli and route via Jyväskylä. In addition, the route between these two in the close proximity of the rail track of Pieksämäki (Map 7.) was mentioned, but after Telia's via-analysis and Google Maps navigator results, the amount of car traffic from Helsinki to Kuopio using this (via Pieksämäki) route was stated to be basically not existing. Hence, a route via Jyväskylä and route via Mikkeli were chosen to represent the route choice options between these municipalities.

When it comes to rail traffic, the rail track used between Helsinki and Kuopio was received from a long-distance train traffic internet service provided by the National Rail Lines (Long distance line map, VR - Kaukoliikenteen reittikartta). The track goes from Helsinki to Lahti, then going via Kouvola to Mikkeli and finally through Pieksämäki to Kuopio (Map 7.).

With the data received from these external spatial data sources (Google and VR), it was possible to build up a comprehensive picture regarding the geographical paths that are used when going from Helsinki to Kuopio. This creates the possibility of detecting mobile network signaling events from the proximity of these paths, and this way assign the information regarding the mode of transportation for the signaling event chains. Basically, if signaling events are gathered from Jyväskylä or from the proximity of highway 5 after Mikkeli, the used mode is most probably a rubber tire vehicle. Then, if the signaling events are gathered from Pieksämäki rail track, the used mode is most probably a train. The observation points that are used for the case Helsinki – Kuopio are marked with red circles in the Map 7. (for more about via-analysis, see section 3.1.7):



Map 7. The via-locations used for the OD-pair Helsinki – Kuopio for distinguishing road and rail traffic.

When it comes to air travel, the ODM via-analysis is not usable, as no signaling events are created during the travel. Hence, air travel will be distinguished from other modes by utilizing the travel duration information. If the distance travelled is long enough (which it often is when travelling by plane), the air traveling can be quite easily distinguished from other modes based on its superior speed. The flight time from Helsinki to Kuopio is one hour, so a trip from an origin in Helsinki region to a destination in Kuopio by using a plane should take no more than 3 – 4 hours in total. However, by land it is not possible to make this travel below 4 hours, which makes it reasonably reliable to assume that trips with durations of 1 – 4 hours are made by plane:

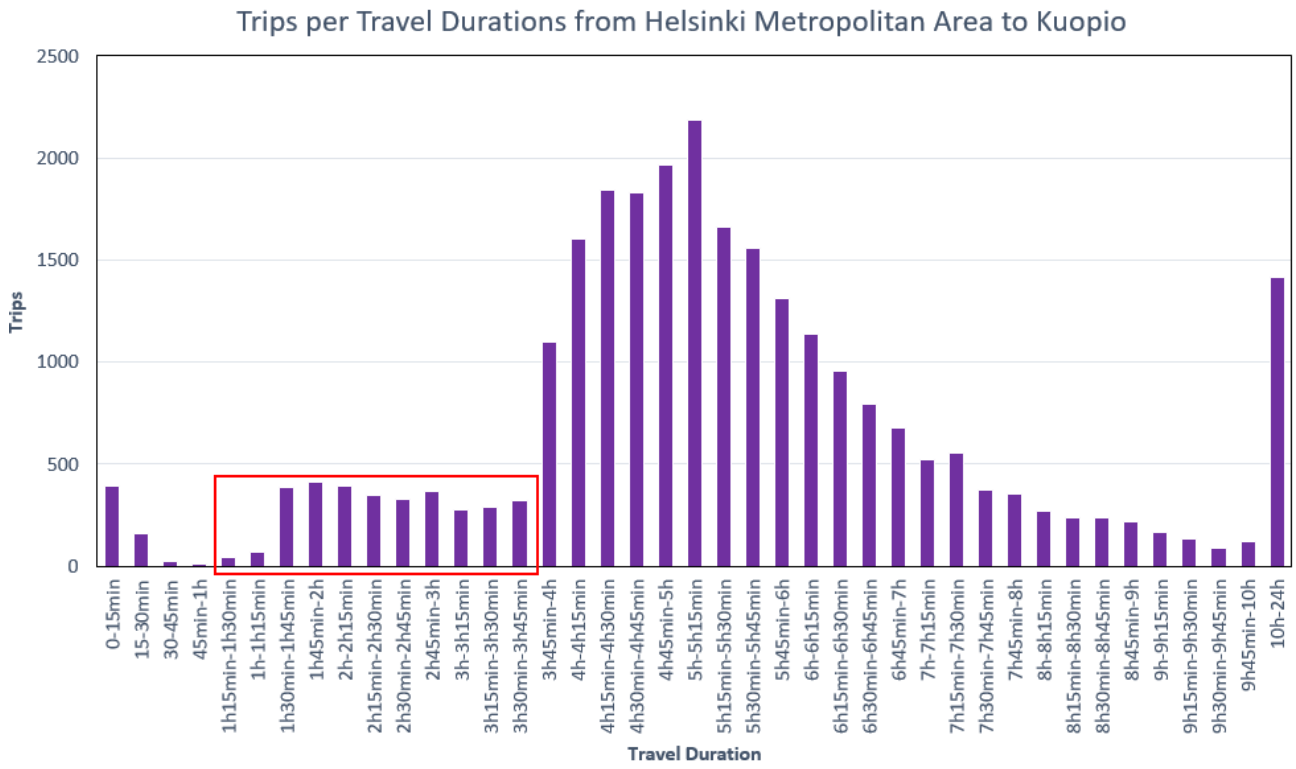


Figure 20. Trips from Helsinki (Helsinki, Vantaa, Espoo) to Kuopio, classified into 15-minute travel duration bins. The air travels are distinguished from other trips based on the assumption, that all air travels are longer than 1 hour but shorter than 4 hours. The air travels are circled with a red square.

As can be seen from the figure 20., Telia's ODM analysis has produced trips from Helsinki to Kuopio that have lasted less than 1 hour. This is a clear indication of noise that has got through the noise filtering algorithms. It is certainly not possible that 450 people have travelled from Helsinki to Kuopio in less than 15 minutes. This is being noticed in the next phases by removing these trips from the Telia's total trip count from Helsinki to Kuopio. As the total trip count of Telia's travel duration ODM analysis matches with Telia's ODM analysis without travel durations included, it can be assumed that every ODM analysis includes a little amount of noise.

After the combination of the via-analysis and the travel duration analysis, we end up having a trip count for road traffic, rail traffic and air traffic. In total, Telia's mobile network data gives a trip count of 26589 (short travel duration noise filtered out) trips from Helsinki to Kuopio during the April of 2019. With the proposed method for modal split, we end up being able to assign a mode for 24 082 trips. Hence, roughly 91 % of the trips are assigned for a mode. This result is compared to the results of National Travel survey 2016 as follows:

Table 10. Comparison between NTS and Telia's mobile network data results regarding trips from Helsinki to Kuopio and modes used on the way. The resulting trip counts are given in the left side of the matrix, next to them are the absolute difference counts (NTS - Telia), and in the right the ratios (NTS / Telia).

	NTS	Telia	Abs. Dif.	Ratio
Road	17838,4	14593,6	3244,83	1,22235
Rail	6681,12	5760,69	920,438	1,15978
Flight	4515,41	3248,07	1267,34	1,39018
Unassigned	—	2956,31	—	—
Total	29035	26558,7	2476,3	1,09324

As can be seen from the Table 10., the total trip ratio between the NTS (yearly trips between Helsinki and Kuopio divided to 12 for a monthly result and then divided by 2 for a directional result) and Telia's mobile network data is very consistent: 1,09. When it comes to the modal split, the relational modal split is almost a perfect match (Figure 21.), even though Telia results with little bit less counts regarding all the modes. Obvious reason for this is that Telia's trip count result was lower already as in total, but also because 9 % of the trips could not been assigned for a mode. The results are presented relatively and absolutely as follows:

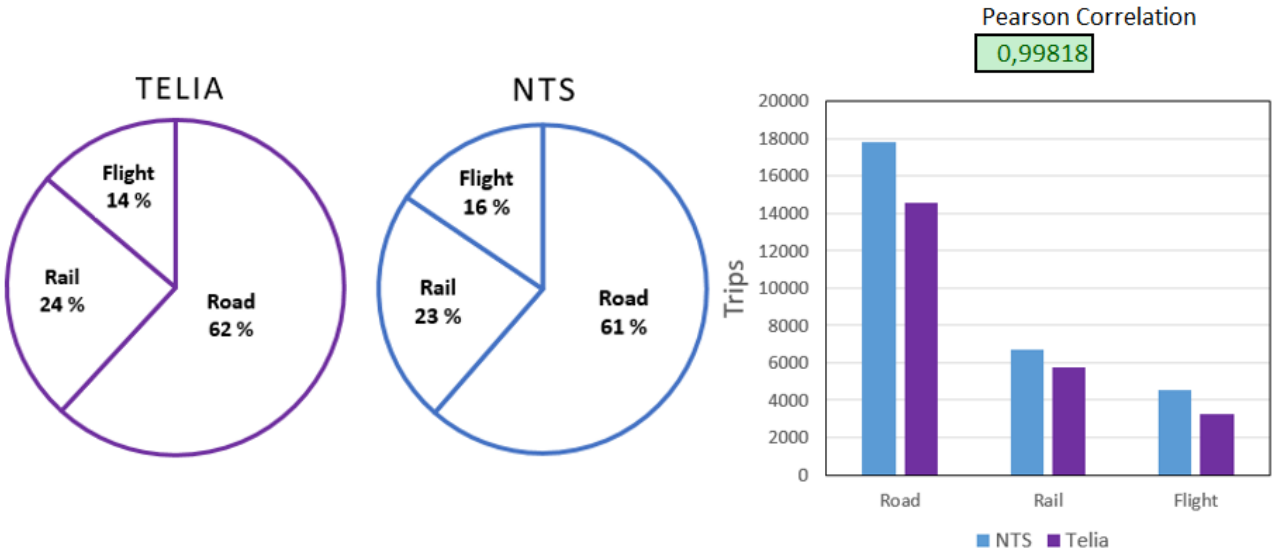


Figure 21. The relational modal split between the data sources (left) and absolute trip counts per transport modes between the data sources (right).

Like assumed, the highway trip assignment was strongly focused on the highway 5 via Mikkeli. 94,6 % of the total highway trips were performed via Mikkeli and only 5,6 % of the travelers chose the route via Jyväskylä. Regarding the fact that Google Navigator recommends the route via Mikkeli exclusively, it might be that Google has surprisingly powerful effect on the route decisions, as the route via Jyväskylä is only few minutes longer in duration and actually 6 kilometers shorter in distance.

5.4.1.2 Case Helsinki – Kuopio, Problems

Major complication regarding modal split from Telia's mobile network data considers the inability to stop long trips. As a trip from Helsinki to Kuopio is several hundred kilometers, a traveler is allowed for relatively long breaks during the trip without stopping it (for trip generation, see section 2.7 and 3.1.2). For example, if a person flies from Helsinki to Kuopio, then walks out from the plain into a taxi, takes ride to home, drops his/her luggage and heads out to the downtown, the trip continues all along. Maybe the person continues by going into a coffeehouse, then walks to a clothing store and maybe even goes for a jog in the evening after spending time in the center. Hence, no activity is detected as the person does not stop moving for over an hour or so, and the trip is continued even though the long-distance travel is already made and ended. This can be clearly seen from the travel duration chart, Figure 20.; there are almost 1500 trips that have lasted over 10 hours.

The complication of too long trips is especially challenging regarding the air traffic separation, as this is completely based on the travel durations. In other words, if someone flies between the municipalities and then continues his/her movement within the city, the proposed air traffic detection fails to capture this trip, even though it is captured to the total trip count. It is possible to adjust the breaking parameter of the trip generation to decrease the allowed breaking time, for example to fixed 30 minutes, but this would then cut out all the road travelers who stop to eat on the way when driving.

In theory, it would be possible to rerun the travel duration analysis with a fixed breaking value of 30 minutes. Then it could be reviewed that how much this would increase the amount of air traffic and simultaneously ignore the decrease in the road traffic that this would cause. Then this increase of air traffic could be summed up with the original results that were generated with the distant dependent breaking value. This would probably increase the amount of air traffic and this way decrease the number of trips with unassigned modes. However, also persons who travel by land can continue their movements after reaching Kuopio, and this is extremely difficult to fix without losing trips simultaneously from the way by.

The continuum of trips in the destination municipality is not the only possible cause for the shortage of modes assigned. It is also possible that the via-locations are not able to capture all the people that are passing by them. For example, if a person drives from Helsinki to Kuopio in five hours while having his/her phone off, this results as having one 5-hour trip in the total trip count, but this trip not assigned into road traffic or rail traffic. This can happen also if the phone is old or does not have internet access; this results in signaling events occurring only when the person makes or receives a phone call or a SMS-message. These operations might be performed more rarely when driving a car than normally, and for this reason there might be too low number of signaling events on the way by for detecting the geographical path used.

In the start of the thesis work project, Finavia's flight statistics (Finavia, 2019) were considered as usable reference data sources for validating air traffic trip counts. From the Finavia's flight statistics it is possible to see the total domestic passenger count of each airport, and as the only scheduled domestic flight from Kuopio airport is headed to Helsinki-Vantaa airport, the total passenger count of Kuopio airport equals the number of trips between Kuopio and Helsinki-Vantaa. However, when the air traffic trip counts between Helsinki-Vantaa and Kuopio from the mobile network data (3 248) and the National Travel Survey (4 515) were compared to the air traffic trip count from Finavia's flight statistics (8

974) it was found that Finavia's results are significantly higher. The reason for this is that as the mobile network data and the NTS indicate trips that origin and finish in Helsinki, Espoo, Vantaa or Kuopio, the airport is also used by lot of other trips. For example, one might fly from Kuopio to Helsinki-Vantaa and then take a bus to Kerava. This trip would not be counted as an air trip between Kuopio and Helsinki, Vantaa or Espoo in the NTS or in the mobile network analysis, even though it was done via Kuopio airport and Helsinki-Vantaa airport.

One simple solution for dealing with the trips that ended up without a mode assigned would be to sum these trips to the total trip counts with modes based according the relative differences between the modal trip counts (Table 10., Telia's relational modal split). In the case of Helsinki – Kuopio, this would mean that 2476*0,16 trips would be added to air trips, 2476*0,24 trips would be added to rail trips and 2476*0,62 trips would be added to road trips. This method would lead to a perfect match with the NTS. However, it would assume that the modal assignment would fail equally in all of the modal classes, which might not be necessarily true. Especially air trips, which are detected with a different method than the rail and road trips, might have different unassignment ratio than rail and road.

5.4.1 Case Helsinki – Lappeenranta

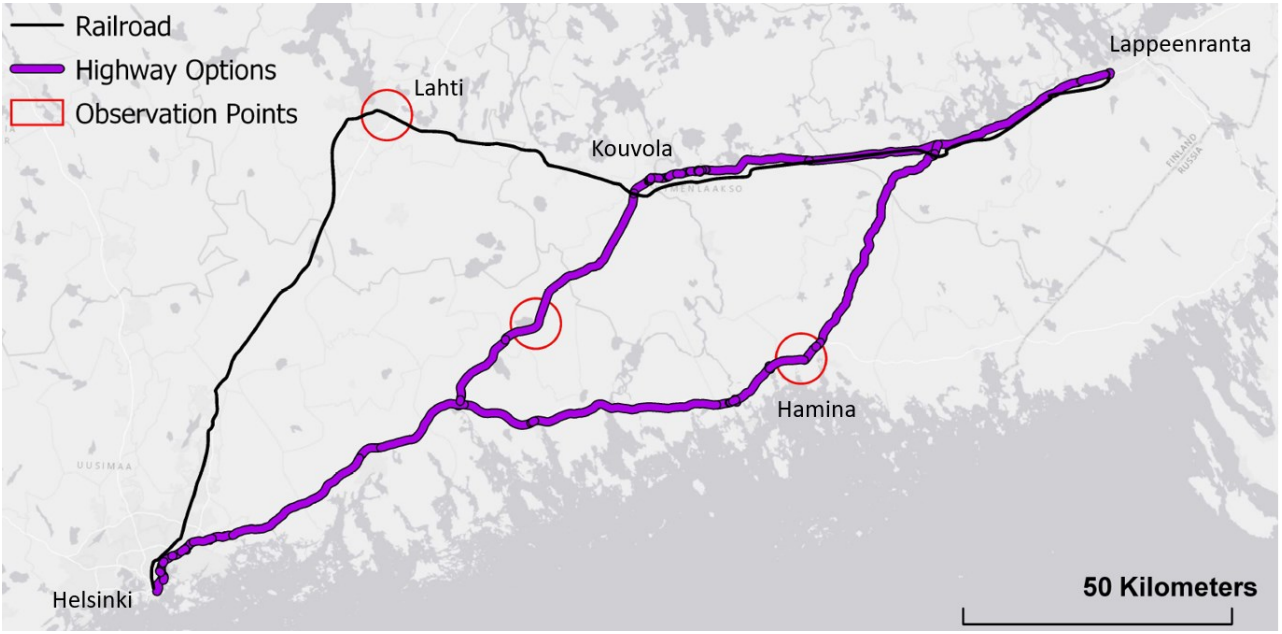
The second case in the route choice and modal split detection analysis was trips from Helsinki (Helsinki, Espoo, Vantaa) to Lappeenranta. Identical temporal and spatial aggregation levels were used than with the Helsinki – Kuopio case, only this time with different destination municipality and geographical paths between the zones. No scheduled flights are operated between Helsinki-Vantaa and Lappeenranta, so air traffic detection was not performed in the case. The OD-pair between Helsinki and Lappeenranta was chosen due its relatively large portion of rail traffic and two different route choice options with almost identical travel times. Telia's market shares (without company subscriptions) that are used for the mobile network data extrapolation are roughly 20 % for Helsinki, Vantaa and Espoo and 25 % for Lappeenranta.

The unextrapolated trip count of the NTS between these zones was 55 trips made by 31 persons. Like with the case Kuopio, it is not a good number of reference trips, but under the circumstances still the best reference available. The limit of 50 trips can be considered to be as a decent amount of unextrapolated reference trips, which is in this case reached.

Again, the analysis was started by going through route choice options given by Google Maps. Even though the time of the day and the day of the week alter the recommendations that Google Map gives, its algorithms recommended exclusively the route via Hamina when driving from Helsinki to Lappeenranta (Map 8.). The Google Map Navigator was ran about ten times during different times of day and week, and every time the route via Hamina was given as the first recommendation by Google, even though the difference to travel time via Kouvola was only 0 – 3 minutes. The route via Kouvola has better overtaking facilities but also more speed cameras along the way, so the set up between the road options was extremely even.

The rail track from Helsinki to Lappeenranta goes from Helsinki to Lahti and then continues via Kouvola to Lappeenranta. As the track takes a detour via Lahti (Map 8.) along the way, it is easy to separate the rail traffic passengers from the highway traffic with a via-point

located in the close proximity of Lahti railway station and highway 12. Then again, the road traffic is detected with via-points located at Lapinjärvi (the route via Kouvola) and Hamina.



Map 8. The Via-locations used for the OD-pair Helsinki – Lappeenranta for distinguishing road and rail traffic.

Results of the via-analysis stated that from the total of 40 491 trips from Helsinki to Lappeenranta in the April of 2019, 20 036 people made their way with a rubber-tire vehicle and 16 915 used a train. 37,4 % of the road trips were performed via Kouvola and 62,6 % via Hamina. 3 541 trips (8,7 % of the total) weren't captured by the via-points placed, so those trips were left without a mode of transport assigned. Unlike with the Kuopio case, also the NTS had in this case a couple of trips that were not assigned for a mode. Reason for this is probably that the answerers of the survey have intentionally or unintentionally left the part regarding the mode of transport unfilled. The results between the datasets are compared as follows (Table 11.):

Table 11. Comparison between NTS and Telia's mobile network data results regarding trips from Helsinki to Lappeenranta and modes used on the way. The resulting trip counts are given in the left side of the matrix, next to them are the absolute difference counts (NTS - Telia), and in the right the ratios (NTS / Telia).

	NTS	Telia	Abs. Dif.	Ratio
Road	15708,52	20035,53	-4327,01	0,784033
Rail	8694,139	16914,73	-8220,59	0,513998
Unassigned	797,1584	3540,845	–	–
Total	25199,82	40491,1	-15291,3	0,622355

Unlike with the Kuopio case, the Lappeenranta case was not a perfect match regarding the total trip counts between the NTS and the mobile network data (Table 11.). Like seen from the trip count validation phase of the Results-section (See great regional ODM, section 5.1), the most probable reason for this was the lack of unextrapolated trips in the NTS. Also,

compared to the Kuopio case, Lappeenranta case had lower answerer extrapolation factors in the NTS than Kuopio. Even though both OD-pairs had roughly 50 unextrapolated trips in the survey, these 50 trips were extrapolated with an average extrapolation factor of 798 in the Kuopio case, whereas in the Lappeenranta case the average extrapolation factor was only 628. When dealing with such a low number of unextrapolated trips, even small changes in extrapolation factors might lead to remarkable differences. For example, with the extrapolation factors of case Kuopio, the case Lappeenranta would have resulted with 7 thousand monthly trips more in total.

However, the total trip count was not the only difference between the results, as the NTS reported significantly lower relative portion of rail trips than the mobile network data. This is more clearly demonstrated with the following charts:

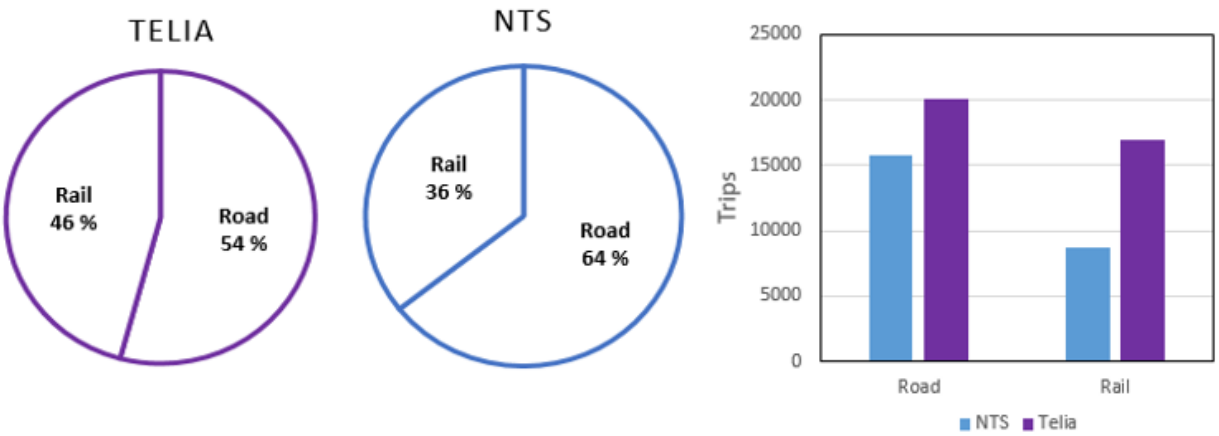


Figure 22. The relational modal split between the data sources (left) and absolute trip counts per transport modes between the data sources (right).

A finding that promotes the assumption of mobile network data being more capable of generating reliable modal split for the OD-pair Helsinki – Lappeenranta than the National Travel Survey 2016, is that the great portion of rail traffic was found not only with the via-analysis but also with an independent trip duration analysis. In the Lappeenranta case, the travel duration feature was not needed for air traffic separation, but it was still generated to acquire information regarding route and rail traffic durations. From the trip duration chart below (Figure 23.), we can see how the trip duration class of "2 h – 2 h 15 min" stands out significantly from the other classes:

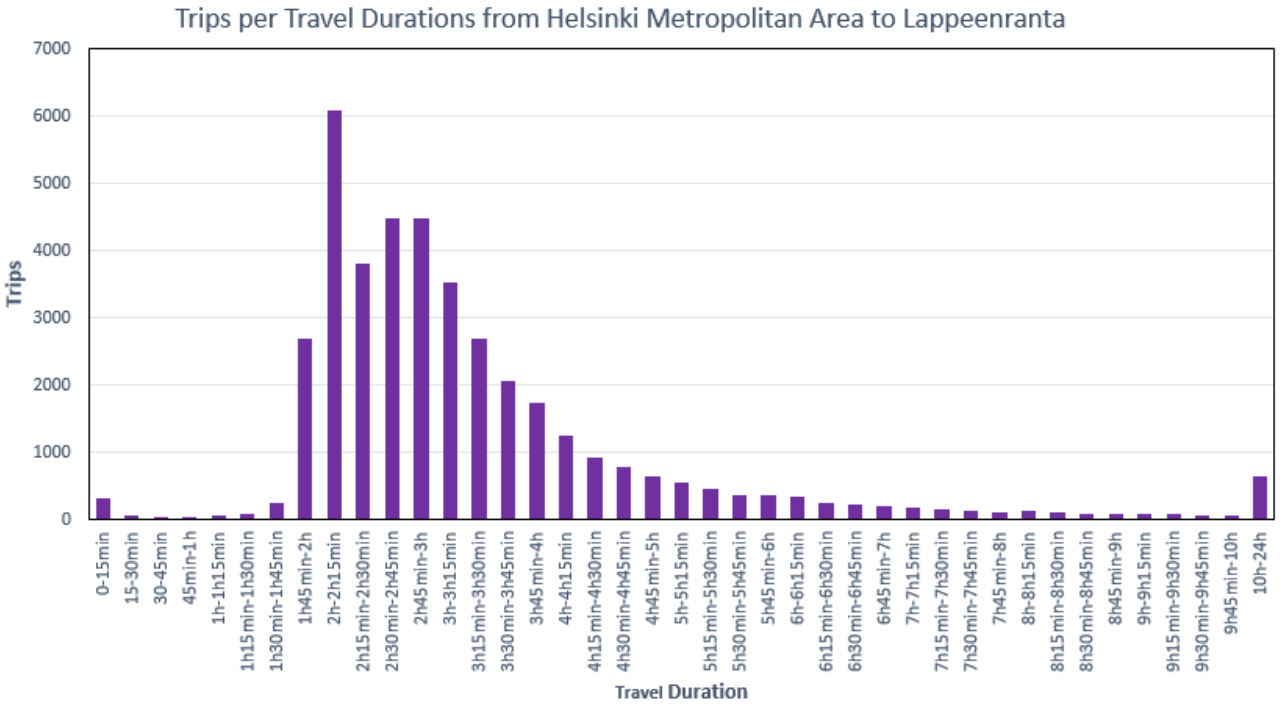


Figure 23. Trips from Helsinki (Helsinki, Vantaa, Espoo) to Lappeenranta, classified into 15-minute travel duration bins.

Regarding the travel duration bin "1 h 45 min – 2 h", it is impossible to perform the trip with a rubber-tire vehicle. The bin "2 h – 2h 15min" is then again theoretically possible if strong speeding is performed, but in practice, not usual. This means that these bins are exclusively rail trips. The bins "2 h 15 min – 2 h 30 min" and "2 h 30 min – 2 h 45 min" include both rail and road trips, and the bins from there forward include mostly road trips. The great portion of exclusive rail trips promotes the results of the via-analysis; rail trips presenting half of the trips between Helsinki and Lappeenranta in total. Like with the Kuopio case, some noise got through the filter algorithms (the trips shorter than 15 minutes), and some trips were continued after reaching Lappeenranta, resulting as trips longer than few hours.

5.4.2 Limitations

As demonstrated with Kuopio and Lappeenranta cases, the proposed method for generating the route choice and modal split results with mobile network data demands independent observations regarding all the OD-pairs that are wanted to be studied. One needs to find out the possible route options, the differing paths for rail tracks and the presumable travel durations for all the modes already in advance to be able measure them with this technique. If one of the mentioned aspects cannot be discovered before the analysis, the analysis cannot be performed.

The most significant limitations are related to the geographical path of the rail track and to the duration of the flight time. For example, if origin X and destination Y are roughly 200 – 300 kilometers from each other and the connecting rail track (which operates relatively slow) goes side by side with the connecting highway all the way, a reliable modal split detection is most probably not possible. This is because the travel duration for a car, a train and for an airplane is roughly the same, about 2 hours. In addition, no differing geographical paths can

be found from the way to distinguish rail and road paths. Theoretically, in this case air traffic could be detected by measuring the sum of rail and road traffic with a via-points on the ground and the difference between this sum and the total trip count.

In addition, the proposed method assumes that all of the highway traffic uses the most optimal routes for the trip. In real life, this is not necessarily true. One might take a detour to pick up a friend from another city or another might take a bike and cycle the whole way using a beautiful coastal cycling path. Even though these trips are captured to the total trip count, the detection of transport mode is difficult.

International traffic creates limitations for the comparability of mobile network data and the National Travel Survey by classifying same stages of trips differently. This happens especially when people from rural areas travel to Helsinki-Vantaa airport to take international flights. In the NTS, this kind of trip is classified as a trip from the origin municipality to the international destination, the main mode of transport being an airplane. In mobile network data however, this is classified as a road or rail trip from the origin municipality to Helsinki-Vantaa. The international flight can be seen after that, as a separate air trip from the airport to the border of Finland. Even though the portion of this kind of trips might not be significant, this is something that must be considered in the comparisons and validations. For example, the case Helsinki – Lappeenranta most probably includes some rail trips between Lappeenranta and Helsinki-Vantaa that are shown in the results of the mobile network data but not in the results of the NTS.

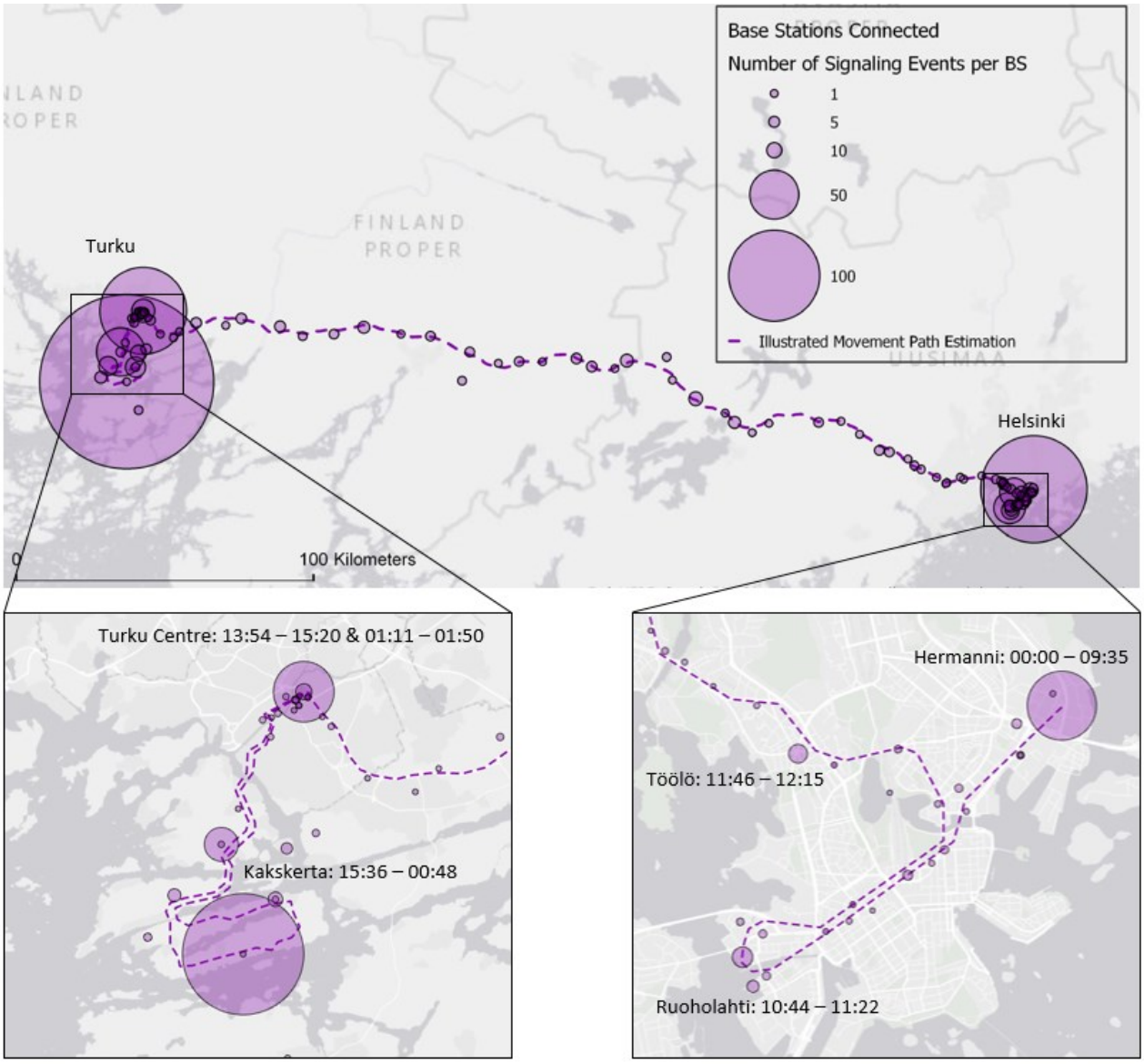
If interested in more accurate modal split than road/rail/air division, mobile network data cannot compete at the moment with surveys. For example, the National Travel Survey 2016 divided the road traffic in Lappeenranta and Kuopio cases to car drivers, car passengers, van drivers, van passengers, scheduled bus lines and unscheduled bus lines. In total, the NTS 2016 has as many as 36 different modal options, including for example husky rides, kick sleds, paragliders, golf cars, rubber dinghies and fire trucks. These kinds of modes are not going to be seen in the modal split of mobile network data at least for a while.

5.5 Whitelisted data analysis

As been discussed throughout the thesis work, the reference data sources for long-distance origin-destination trips and their attributes in Finland are not the best possible. One way to validate OD-matrices generated with mobile network data without external reference data sources is to take a look into the raw signaling event data, from which the trips and activities are generated. For privacy reasons, this is obviously not possible regarding the subscriptions in general, but for business development purposes Telia Finland has whitelisted three of its employees' mobile phone subscriptions just for the purpose with the consent from these employees. It means, that the raw data from these three phones is always available for analysis and validation.

The attributes of the raw signaling event data are given accurately in sections 2.4 and 2.5. Basically it means that every network connection is mapped geographically based on the base station coordinates which received the connections. Hence, we can see how many times a mobile phone was connected to each base station and when did the connections happen. The result is a map showing chronologically every connection and their locations during a day (Map 9.).

For the analysis case, a timeframe of 24.8.2019 00:00 – 25.8.2019 02:00 was chosen. For this timeframe it was known that one of the whitelisted phones travelled from Helsinki to Turku. The signaling events were not aggregated or assigned temporally or spatially to any other unit or class; only base station locations were used spatially, and the actual connection times presented as they were in seconds. However, seconds are rounded into minutes in the resulting Map 3.:



Map 9. Raw signaling event data visualization during 24.8.2019 for one whitelisted mobile phone subscription. Base stations that are connected are presented with violet dots; dot size presents the amount of connections regarding that particular base station. An illustration of the mobile phone movement path is drawn chronologically from the morning to the evening, based on the base station connections during the day. Activities that can be easily separated from trips are labeled with the location name and the time of the stay.

As can be seen from the Map 3., trips and activities can quite easily be detected from the raw signaling event data. In this particular case, 1031 signaling events in total were connected to

107 different base stations during the day. This means that a signaling event occurred approximately once in a minute, whereas one base station was being connected 13,5 minutes in average. Longest activities occurred in Hermanni (the night between 23.8. and 24.8.) and in Kakskerta Turku (15:36 – 00:48). Both of these activities created over 100 signaling events. The base stations along the way at Turku highway received 2,48 connections per base station in average.

Regarding the results of the whitelisted data analysis there is absolutely no question whether mobile network data can detect a trip from a municipality to another. The spatial accuracy of the results is so accurate, that also the shorter stops along the way, Ruoholahti, Töölö and Turku centre, are captured with a variance of about 0 – 200 meters to the real location visited. And as BSE:s are used in Telia's trip generation algorithms for the positioning instead of base station locations, even more accurate spatial results are formed when OD-analyses are generated.

6. Discussion

6.1 This study in relation to earlier work

By starting from the properties of the mobile network data itself, the great majority of the analyses in the related work is still performed with a relatively small set of Call Detail Records (CDR). Regularly, this means a timeframe of month or two, with a population count of some hundred thousands (For exception, see Zagatti et al., (2018) with 2 billion rows of signaling event data). Using sparse-in-time featured CDR data results in missing accuracy in the analyses (Kujala et al., 2016) and makes modal split detection extremely difficult (Huang et al., 2019). In addition, it makes small scale spatial aggregation of OD-matrices unreliable, as short trips are more difficult to detect with lower number of signaling events (Alexander et al., 2015). What can also be seen from earlier studies, is that Best-Server-Estimates (BSE:s) are not often used (Bonnel et al., 2018; Gonzales et al., 2008; Zagatti et al., 2018) in the geographical assignment of trips and activities, but Voronoi tesselation is applied instead. The reason for this is probably that mobile network operators are not eager to provide their real cell boundaries for researchers. Referring to Telia's BSE:s, Voronoi tesselation might be surprisingly weak method for estimating mobile network cell boundaries, as LTE (4G) and GSM antennas might provide completely different sized cells.

In the respect of the limitations of the earlier studies mentioned above, the thesis work analyses can be said to be performed with world-class mobile network data. With the combination of passive and active signaling events, utilization of BSE:s and the comprehensive distribution of base stations, Telia's raw signaling event data provides accurate movement paths in time and space without biases, at least when it comes to its own subscribers (See section 5.5, whitelisted analysis). In this sense, the whitelisted data analysis promotes the findings of Chen et al., (2014), where the spatial accuracy of passive signaling events are detected to be usually within 100 meters of the actual location visited by the user.

When it comes to actual origin-destination analyses with mobile network data, similar results were obtained in the thesis work project compared to earlier work. Whereas Bonnel et al., (2018) and Graells-Garrido & Saez-Trumper, (2016) came up with correlations of around 0,9 between OD-matrices by household travel surveys and mobile network data, the correlations achieved in the thesis work were even higher. However, different things were analyzed from the datasets, as the thesis work focused on the trip counts, whereas Graells-Garrido & Saez-Trumper (2016) incorporated trip attributes such as starting time, travel duration and travel distances to the correlation analyses. In addition, studies by Bonnel et al., (2018) and Graells-Garrido & Saez-Trumper, (2016) focused on short trips, whilst the thesis work analyzed long trips. Hence, the difference between correlation values makes sense, as long trips are considered to be easier to detect reliably with mobile network data than short trips.

Referring to Bekhor et al., (2013) and Alexander et al., (2015), national household travel surveys tend to underestimate OD trip counts regarding trips that are not regular. For example, Alexander et al., (2015) points out that the trips that are occuring late at evenings may not be reported to surveys, whereas Bekhar et al., (2013) demonstrates that long trips are under-presented in the regular survey set ups. These assumptions strongly promote the findings of the thesis work. By highlighting the thesis work observations regarding summer-peak inconsistencies (see section 5.1.3), trip count inconsistencies in great regional ODM

(see section 5.1.1) and rail traffic ratio differences in case Helsinki – Lappeenranta (see section 5.4.2), this assumption seems most probable. After all, the primary aim of the National Travel Survey is to present regular travel patterns, not marginal or irregular travel patterns.

Zagatti et al., (2018) demonstrated the usefulness of mobile network data in data poor environments. The thesis work results promote this finding, as Finnish long-distance transportation between rural municipalities can be considered as a data poor environment. Characteristics of Finnish demography and geography shape up the national transportation in a way that regions with low inhabitant counts are connected to other low inhabitant regions with long distances. Out of 235 world countries, Finland stands as 200th in population density (United Nations, 2019). Even without Lapland, Finland stands still only as 185th in population density. This results in difficulties to conduct reliable origin-destination matrices with travel surveys, especially if either origin or destination is not a large municipality or region (see section 4.2.3, Regional ODM). Compared to other related studies in which mobile network data has been validated against household travel surveys, the inability of survey data being used to produce OD-matrices has not stood out that significantly. Based on this, it might be that Finland has even better circumstances to benefit from the usage of mobile network data in long-distance transportation planning than most of the other western countries.

When it comes to detection of modal split, the results of the thesis work correlate with other related work. Studies by Garcia et al., (2016) and Smoreda et al., (2013) distinguished between road, rail and air in inter-city OD-trips, with similar rule-based methodology than the one used in the thesis work. Referring to Huang et al., (2019) the detection of main modes (road, rail, air) in long-distance transportation is relatively easy with network-driven mobile netowrk data, which was demonstrated in the thesis work. However, Garcia et al., (2016) were able to detect as much as 98 % of the modes used, whereas the thesis work cases of Helsinki – Kuopio and Helsinki – Lappeenranta resulted in about 90 % of trips assigned to a mode. One reason for this difference might have been that Garcia et al., (2016) used only airport locations for the air traffic detection instead of travel durations. This is an effective method when airports are located in clearly divergent locations from the other modal paths. In the thesis work this method was not applied due the via-point placing based on OSM street nodes instead of individual base stations.

Similarly to the results of Garcia et al., (2016), the methodology and data used in the thesis work seem to be able to not only provide correlating results with the National Travel Survey 2016, but also provide additional insights that cannot be conducted from other data sources. Hence, there are three main channels for mobile network data to stand out from the conventional data:

1. Conventional data sources provide faulty results (see sections 5.1 summer-peak inconstistencies, faulty trip counts in rural OD-pairs and 5.4.2 questionable modal shares).

2. Mobile network data is capable of providing completely new information (see sections 5.4 information on route choice and 5.1.2 temporal trip count changes).

3. Mobile network data is capable of providing correlating results with conventional data sources, with lower costs (see sections 5.1 Great regional ODM, 5.2 Greater Helsinki Municipality ODM, 5.4.1 Case Helsinki – Kuopio).

Referring to most of the earlier studies done in the field, mobile network data is considered as more cost-effective data source than household travel surveys. When it comes to actual euros, division of costs and their comparison between data sources is however not simple. The factor that makes it challenging, is the definition of the components and use cases in household travel surveys that could be replaced with mobile network data, and then again the definition of the components that should be kept. In the case Finland, the delivery of the National Travel Survey 2016 cost in total over 1 million euros, in which the portion of the national survey section was roughly a half (Finnish Transport Agency, 2013). Referring to the results of the thesis work, seems unlikely that the costs of the national travel surveys in Finland could not be decreased by utilizing mobile network data. This is of course dependent on the information that is wanted to be acquired with the trasnportation studies.

6.2 Limitations of the study

There are three main limitation categories in the thesis work study:

1. The mobile network data being tied into one operator and its data processing methodology.
2. Extrapolation methodology.
3. Availability and correctness of reference data sources.

What was naturally known already before starting the thesis work project, mobile network data was available from only one telecom operator, Telia. In addition, raw signaling event data was used only in the whitelisted data analysis (see section 5.5), which meant that all other analyses used already processed data, provided by Telia's data processing algorithms. This creates limitations regarding repeatability of the study, as similar circumstances cannot be created elsewhere. This creates also challenges in recognizing the factors that may have or may have not affected the results positively or negatively. In other words, the "black boxing" of the trip generation, anonymization, extrapolation and aggregation processes hides the factors that may function correctly or may provide faulty results. All we know is that the sum of these sub-processes, the results of the study, seems to be valid and reliable when compared to reference data sources. The question, could some of the sub-processes function even better, is difficult to answer at the moment.

Referring to the sub-process of extrapolation, seems to be that it still needs more development. When extrapolation of mobile network data and household travel surveys are compared, the differences in the qualities of extrapolation are clear. Whereas the National Travel Survey 2016 used four different attributes for the extrapolation (home location, gender, age and level of education), mobile network data has to get along with only CRM-based home location of the subscription or with the location of the mobile phone's first-signal at morning. This is typical methodology in the related work (e.g. Alexander et al., 2015; Bonnel et al., 2018), but its limitations are widely known. When dealing with this kind of a simple extrapolation methodology, some strong assumptions must be made: 1. Gender and age do not affect personal travel behavior, 2. Gender and age do not affect mobile phone usage behavior and 3. Urbanity or rurality of the home location does not affect the representativeness of the observation (see section 2.8, Representativeness). As we know however, these factors do have an effect in the real world, which makes the simple extrapolation methodology inferior to more complicated ones. Fortunately, telecom

operators do have a possibility of combining age and gender information to the extrapolation equations; in the end it is just a matter of data gathering and warehousing.

Another limitation of the current trip extrapolation method used in the thesis work is that it is targeted to the trips themselves and not persons performing the trips. This means that it does not matter "who" performs the trip, as the trip is always extrapolated based on the market share of the origin postal zone and the market share of the destination postal zone. In the National Travel Survey then again, it does not matter where the trip is performed, as it is always extrapolated based on the personal extrapolation factor of the survey participant. The method that is currently used by Telia's algorithms, might lead to biases created by irregular postal zone market shares, for example like in office districts.

Third and final limitation category concerned the reference data sources. Even though it was expected before starting the study, the availability and correctness of reference data sources was surprisingly weak. The National Travel Survey 2016 had too low number of long-distance trips observed, Finavia's flight statistics were too inaccurate, rail statistics of VR were not available and flow information gathered from LAM-data and rail flow counts by Finnish Transport Agency could not be reliably converted into origins and destinations. In addition, route choice data in Finland is practically non-existing. All this resulted in a way that the validity of trip counts from mobile network data was only barely checked, whereas the results regarding modal split and route choice detection were basically just demonstrated, not validated. All in all, it is quite alerting that Finnish long-distance transportation is such a mystery at the moment. It clearly demonstrates the demand of alternative and new data sources.

6.3 Future Research

Referring to the results of the thesis work, seems to be that in addition to further investments in mobile network data validation in Finland, the research emphasis of the field should be shifted from pure data validation towards analyzing the parts of conventional transport research methods that can be replaced with mobile network data. In other words, cost management research regarding Finnish transport data utilization should be performed, in which the best role for mobile network data in transport studies in whole should be searched. This means finding the elements of surveys and traffic census which are more expensive to use than mobile network data, but still cannot bring more value. Strong expertise in addition to data itself are required regarding the processes of transport research and transport data analysis in Finland, in addition to understanding what the needs and processes of the stakeholders of Finnish transportation are. Deep-rooted traditional processes can be difficult to change without clear observations regarding cost-effectiveness, demonstrated in euros.

Regarding the mobile network data quality, future research should be targeted to the extrapolation process. Huang et al., (2019) conducted a research where home location, gender and market-shares of operators were studied to find out how to best represent the population of a certain city with mobile network data. Age information should be definitely added into the equation and the study area should be widened to a whole country. Not only should it be studied what kind of customer base provides the best extrapolation results for the operator, but also with what statistical methodology can operators best extrapolate their different customer bases. With the right customer data and the extrapolation methodology,

mobile network data can be extrapolated significantly more reliably than surveys, as the sample sizes of mobile network data are remarkably larger.

Besides the extrapolation-related limitations, few other development areas for Telia's mobile network data processing can be suggested. Travel duration feature (see section 5.4) showed that impossibly fast and slow trips are created in addition to the real trips. As the too fast trips are clearly noise that gets through the noise filtering algorithms, the too slow trips are then again created due the inability of stopping the long trips. One solution for stopping the long-distance trips after reaching the destination zone would be to apply special parameters regarding stopping trips in the destination zone even if clear activities are not performed. In addition, only 90 % of long-distance trips were assigned to transport modes. Referring to Garcia et al. (2016) and to the results of the thesis work, implementing via-locations to the airports could increase the amount of air trips detected.

Another solution for trying to understand the profile of durationally too long trips would be too include travel length detection algorithm. Referring to the results of the whitelisted data analysis (section 5.5), it should not be too difficult to measure the lenth of a trip, either with a direct distance or by the road network. This would help to estimate the correctness of the results of the travel duration feature; is the trip really very slow or has it actually made a detour or continued its movement in the destination zone.

When it comes to Finnish long-distance transportation reference data, future research should focus on utilizing more comprehensive set of data sources than just the National Travel Survey, which was the main reference data source used in the thesis work. As long as conducting nation-wide accurate OD-matrices is not an objective of the National Travel Survey, it cannot be reliably used to validate other data sources as it may not be valid itself. Additional data sources which were not utilized in their full potential or at all in the thesis work include for example GPS-data, LAM-data and raw signaling event mobile network data. For example, a combination of raw signaling event data and GPS-tracks gathered from a sample size of 50 people could provide very accurate movement paths that could be compared. LAM-data could then again be used in more aggregated level, by deriving origins, destinations and route choices from traffic flow counts (Cascetta, 2009 s.513; Ortuzar, 2011 s.434).

7. Conclusion

This thesis work validated usage of Telia's mobile network data in Finnish long-distance transport modeling. The mobile network dataset included passive and active mobile network data logs from a timeframe of one year. Movements of roughly 1,8 million Finnish people were monitored during that time. After trip generation, geographical mapping, aggregation and anonymization, movements of these people were formatted to origin-destination matrices. Different OD-matrices were generated regarding different timeframes and spatial levels. These origin-destination matrices were then compared to reference data sources in order to find out whether the matrices created with mobile network data matched with the matrices made with reference data sources.

Trip counts generated from mobile network data between distant Finnish great regions were compared to trip counts of the National Travel Survey 2016. High correlation was achieved between zones with high population counts, whereas inconsistencies were found between two rural OD-pairs with low population counts. Similar results were acquired from the comparison between mobile network data and HELMET transport demand model between Greater Helsinki municipalities. High correlations were achieved regarding larger municipalities whereas trips between small municipalities were more inconsistent. Scarcity of real observed trips in the NTS were proportional to inconsistencies between the data sources, which promotes the assumption that the major inconsistencies between the data sources were caused by the falsity of the NTS. However, this assumption was not statistically proven in the thesis work, as the aim of the study was not to prove the NTS wrong, but to use it to validate mobile network data. This objective was achieved between large municipalities and great regions with Pearson Correlation factor of almost 1,0.

In addition to being able to generate coherent trip count results with reference data sources, the thesis work demonstrated additional insights regarding seasonal changes and route choice information, that cannot be acquired from other data sources. Total seasonal travel count changes were validated against LAM-data and reliable results were acquired, which divided Finnish long-distance transportation into three main classes; low level winter, medium level fall and high level summer. Temporal advantages of mobile network data were demonstrated by detecting the inability of the NTS to capture high summer-peaks in trip counts, when Finns massively migrate to their cottages.

The thesis work proposed a rule-based method for detecting modal split between a known origin and a destination with mobile network data. The ability was validated against the NTS with two OD-pair cases. These cases were trips between Helsinki and Kuopio during one month and trips between Helsinki and Lappeenranta during the same month. The results of the analyses showed that if the OD-pairs that are being analyzed include a geographically diverging rail track from the highway options, and if the option of flying is durationally differing from other travel options, the modal split detection between road, rail and air can be done reliably. In both cases the method was able to assign a mode for 90 % of the total trips. The case Helsinki – Kuopio acquired high correlation between data sources, whereas the case Helsinki – Lappeenranta produced differences between mobile network data and the NTS. However, the validity of mobile network data -based modal split detection is difficult to confirm, as the results of the NTS are most probably not valid, like assumed with the trip counts.

In addition to some more validation, future emphasis should be shifted from pure validation into consideration of more productional utilizations of mobile network data in Finnish

transportation. In addition to using mobile network data in transport demand modelling, which is already taking place, it would be beneficial to deeply understand which elements of the current travel survey processes could be replaced with utilization of mobile network data in order to acquire better results and cost savings. At least when it comes to origin-destination modeling, the results of this thesis work demonstrate the superiority of mobile network data being used to create OD-matrices also in rural areas with low population counts. If this is something that is wanted to be achieved in the whole national level, utilization of mobile network data as one of the main data sources should be strongly considered.

References

Ahas, R., Aasa, A., Roose, A., Mark, Ü., & Silm, S. (2008). Evaluating passive mobile positioning data for tourism surveys: An Estonian case study. *Tourism Management*, *29*(3), 469-486.

Alexander, L., Jiang, S., Murga, M., & González, M. C. (2015). Origin–destination trips by purpose and time of day inferred from mobile phone data. *Transportation research part c: emerging technologies*, *58*, 240-250.

Asakura, Y. and Hato, E., (2004). Tracking survey for individual travel behaviour using mobile communication instruments. Transportation Research Part C: Emerging Technologies, 12 (3–4), 273–291.

Asgari, F., Gauthier, V., and Becker, M. (2013). A survey on Human Mobility and its applications. arXiv preprint arXiv:1307.0814.

Aubrecht, C., Steinnocher, K., Hollaus, M., & Wagner, W. (2009). Integrating earth observation and GIScience for high resolution spatial and functional modeling of urban land use. *Computers, Environment and Urban Systems*, *33*(1), 15-25.

Bachir, D., Khodabandelou, G., Gauthier, V., El Yacoubi, M., & Puchinger, J. (2019). Inferring dynamic origin-destination flows by transport mode using mobile phone data. *Transportation Research Part C: Emerging Technologies*, *101*, 254-275.

Bakillah, M., Liang, S., Mobasheri, A., Jokar Arsanjani, J., & Zipf, A. (2014). Fine-resolution population mapping using OpenStreetMap points-of-interest. *International Journal of Geographical Information Science*, *28*(9), 1940-1963.

Benesty, J., Chen, J., Huang, Y., & Cohen, I. (2009). Pearson correlation coefficient. In *Noise reduction in speech processing* (pp. 1-4). Springer, Berlin, Heidelberg.

Bekhor, S., Cohen, Y., & Solomon, C. (2013). Evaluating long-distance travel patterns in Israel by tracking cellular phone positions. *Journal of Advanced Transportation*, *47*(4), 435-446.

Beresford, A. R., & Stajano, F. (2003). Location privacy in pervasive computing. *IEEE Pervasive computing*, (1), 46-55.

Biljecki, F., Ohori, K. A., Ledoux, H., Peters, R., & Stoter, J. (2016). Population estimation using a 3D city model: A multi-scale country-wide study in the Netherlands. *PloS one*, *11*(6), e0156808.

Blumenstock, J., Cadamuro, G., and On, R., (2015). Predicting poverty and wealth from mobile phone metadata. Science, 350 (6264), 1073–1076. doi:10.1126/science.aac4420

Bonnel, P., Fekih, M., & Smoreda, Z. (2018). Origin-Destination estimation using mobile network probe data. *Transportation Research Procedia*, *32*, 69-81.

Bolla R, Davoli F (2000), Road Traffic Estimation from Location Tracking Data in the Mobile Cellular Network, Proc. IEEE WCNC, Chicago, USA

Candia, J., González, M. C., Wang, P., Schoenharl, T., Madey, G., and Barabási, A. L. (2008). Uncovering individual and collective human dynamics from mobile phone records. Journal of Physics A: Mathematical and Theoretical, 41(22), 224015.

Calabrese, F. (2011). Estimating Origin-Destination Flows Using Mobile Phone Location Data. IEEE Pervasive Computing, 10(4), pp. 36-44. doi:10.1109/MPRV.2011.41

Calabrese, F., Diao, M., Di Lorenzo, G., Ferreira Jr, J., & Ratti, C. (2013). Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. *Transportation research part C: emerging technologies*, *26*, 301-313.

Calabrese, F. (2015). Urban Sensing Using Mobile Phone Network Data: A Survey of Research. ACM Computing Surveys (CSUR), 47(2), pp. 1-20. doi:10.1145/2655691

Cascetta, E. (2009). *Transportation systems analysis: Models and applications* (2nd ed.). New York: Springer.

Chen, C., Bian, L., & Ma, J. (2014). From traces to trajectories: How well can we guess activity locations from mobile phone traces?. *Transportation Research Part C: Emerging Technologies*, *46*, 326-337.

Cheng, P., Qiu, Z., & Ran, B. (2006, September). Particle filter based traffic state estimation using cell phone network data. In *2006 IEEE Intelligent Transportation Systems Conference* (pp. 1047-1052). IEEE.

Demissie, M. G., de Almeida Correia, G. H., and Bento, C. (2013). Intelligent road traffic status detection system through cellular networks handover information: An exploratory study. Transportation Research Part C: Emerging Technologies, 32, 76-88.

De Montjoye, Y. A., Hidalgo, C. A., Verleysen, M., & Blondel, V. D. (2013). Unique in the crowd: The privacy bounds of human mobility. *Scientific reports*, *3*, 1376.

Deville, P., Linard, C., Martin, S., Gilbert, M., Stevens, F. R., Gaughan, A. E., ... & Tatem, A. J. (2014). Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences*, *111*(45), 15888-15893.

Dobson, J. E., Bright, E. A., Coleman, P. R., Durfee, R. C., & Worley, B. A. (2000). LandScan: a global population database for estimating populations at risk. *Photogrammetric engineering and remote sensing*, *66*(7), 849-857.

Dong, H., Wu, M., Ding, X., Chu, L., Jia, L., Qin, Y., & Zhou, X. (2015). Traffic zone division based on big data from mobile phone base stations. *Transportation Research Part C: Emerging Technologies*, *58*, 278-291.

Duckham M, Kulik L (2005). A formal model of obfuscation and negotiation for location privacy. In: 3rd international conference on pervasive computing (Pervasive 2005). Springer, Munich, Germany, pp 152–170

Sara Fayaz & Sara Sarrafian (2014). Location service for wireless network using improved RSS-based cellular localisation, International Journal of Electronics, 101:6,763-778, DOI: 10.1080/00207217.2013.794492

Finavia (2019). Passengers by Airports. Retrieved from: https://www.finavia.fi/sites/default/files/documents/Passengers%20by%20Airport-fi_19.pdf

Finnish Transport Agency (2013). Esiselvitys, Vatakunnallinen henkilöliikennetutkimus 2015 – 2016. Liikenneviraston tutkimuksia ja selvityksiä. Retrieved from:

https://julkaisut.vayla.fi/pdf3/lts_2013-53_valtakunnallinen_henkiloliikennetutkimus_web.pdf

Finnish Transport Agency (2016). Liikenneviraston liikennelaskentajärjestelmä – Päivitetty järjestelmänkuvaus (in Finnish only). Liikenneviraston tutkimuksia ja selvityksiä 36/2016. Retrieved from https://julkaisut.liikennevirasto.fi/pdf8/lts_2016-36_liikenneviraston_liikennelaskentajarjestelma_web.pdf

Finnish Transport Agency (2018a). National Travel Survey 2016 (in Finnish only). Liikenneviraston tilastoja 1/2018. Retrieved from: https://julkaisut.vayla.fi/pdf8/lti_2018-01_henkiloliikennetutkimus_2016_web.pdf

Finnish Transport Agency (2018b). National Travel Survey 2016 – technical report (in Finnish only). Liikenneviraston tutkimuksia ja selvityksiä 14/2018. Retrieved from: https://julkaisut.liikennevirasto.fi/pdf8/lts_2018-14_henkiloliikennetutkimus_tekninen_web.pdf

Finnish Transport Agency (2019a). LAM-tiedot (in Finnish only). Retrieved from: https://vayla.fi/avoindata/tieverkko/lam-tiedot#.XX9T-SgzZPY

Finnish Transport Agency (2019b). Tieliikenteen kehitys (in Finnish only). Retrieved from: https://vayla.fi/tilastot/tietilastot/tieliikenteen-kehitys#.XV56sOgzZPY

García, P., Herranz, R., & Javier, J. (2016). Big data analytics for a passenger-centric air traffic management system. In *Presented at the 6th SESAR Innovation Days, Delft, Netherlands*.

Gedik. B. and Liu. L. (2008). Protecting location privacy with personalized k-anonymity: Architecture and algorithms. IEEE Transactionson Mobile Computing 7,1 (Jan.2008), 1–18. DOI: http://dx.doi.org/10.1109/ TMC.2007.1062

Girardin. F., A. Vaccari, A. Gerber and C. Ratti. (2009). Quantifying urban attractiveness from the distribution and density of digital footprints. Journal of Spatial Data Infrastructure Research 4 (2009), 175–200.

González, M. C., Hidalgo, C. A., and Barabási, A. L. (2008). Understanding individual human mobility patterns, Nature, 453, 779–782.

González F, Melo-Riquelme C, de Grange L (2016) A combined destination and route choice model for a bicycle sharing system. Transportation 43(3):407–423. https://doi.org/10.1007/s11116-015-9581-6

Graells-Garrido, E., Caro, D., & Parra, D. (2018). Inferring modes of transportation using mobile phone data. *EPJ Data Science*, *7*(1), 49.

Graells-Garrido, E., & Saez-Trumper, D. (2016, May). A day of your days: estimating individual daily journeys using mobile data to understand urban flow. In *Proceedings of the second international conference on IoT in urban space* (pp. 1-7). ACM.

Greger, K., (2015). Spatio-temporal building population estimation for highly urbanized areas using GIS: spatio-temporal building population estimation. Transactions in GIS, 19 (1), 129–150. doi:10.1111/tgis.2015.19.issue-1

Hariharan R., Toyama K. (2004) Project Lachesis: Parsing and Modeling Location Histories. In: Egenhofer M.J., Freksa C., Miller H.J. (eds) Geographic Information

Science. GIScience 2004. Lecture Notes in Computer Science, vol 3234. Springer, Berlin, Heidelberg

Heanue, K. E., & Pyers, C. E. (1966). A comparative evaluation of trip distribution procedures. *Highway Research Record*, *114*, 20-50.

Horanont, T., Phiboonbanakit, T., & Phithakkitnukoon, S. (2018). Resembling Population Density Distribution with Massive Mobile Phone Data. *Data Science Journal*, *17*.

Horn, C., Klampfl, S., Cik, M., Reiter, T. (2014). Detecting outliers in cell phone data: correcting trajectories to improve traffic modeling. In: Transportation Research Board 93rd Annual Meeting, 14-3690

HSL (2014). Helsingin seudun työssäkäyntialueen liikenne-ennustejärjestelmän kysyntämallit 2014 (in Finnish only). Retrieved from: https://www.hsl.fi/sites/default/files/21_2016_kysyntamalliraportti.pdf

Huang, H., Cheng, Y., & Weibel, R. (2019). Transport mode detection based on mobile phone network data: A systematic review. *Transportation Research Part C: Emerging Technologies*.

Iqbal, M. S., Choudhury, C. F., Wang, P., & González, M. C. (2014). Development of origin–destination matrices using mobile phone call data. *Transportation Research Part C: Emerging Technologies*, *40*, 63-74.

Isaacman. S., R. Becker, R. C´aceres, S. Kobourov, M. Martonosi, J. Rowland, and A. Varshavsky. (2011). Identifying important places in people's lives from cellular network data. In Proceedings of the 9th International Conference on Pervasive Computing (Pervasive'11). Springer-Verlag, Berlin, Heidelberg, 133–151.

Jahangiri, A., & Rakha, H. A. (2015). Applying machine learning techniques to transportation mode recognition using mobile phone sensor data. *IEEE transactions on intelligent transportation systems*, *16*(5), 2406-2417.

Janecek, A., Valerio, D., Hummel, K. A., Ricciato, F., & Hlavacs, H. (2015). The cellular network as a sensor: From mobile phone data to real-time road traffic monitoring. *IEEE transactions on intelligent transportation systems*, *16*(5), 2551-2572.

Jiang. S., G. A. Fiore, Y. Yang, J. Ferreira, Jr., E. Frazzoli, and M. C. Gonz´alez. (2013). A review of urban computing for mobile phone traces: Current methods, challenges and opportunities. In Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing (UrbComp'13). ACM, New York, NY. DOI:http://dx.doi.org/10.1145/2505821.2505828

Jiaqi, K. C. (2018). *Transport Mode Detection Using Cellular Signaling Data (Case Study of Graz and Vienna, Austria)* (Doctoral dissertation, Geographisches Institut der Universität Zürich).

Järv, O., Tenkanen, H., & Toivonen, T. (2017). Enhancing spatial accuracy of mobile phone data using multi-temporal dasymetric interpolation. *International Journal of Geographical Information Science*, *31*(8), 1630-1651.

Järv, O., Müürisepp, K., Ahas, R., Derudder, B., & Witlox, F. (2015). Ethnic differences in activity spaces as a characteristic of segregation: A study based on mobile phone usage in Tallinn, Estonia. *Urban Studies*, *52*(14), 2680-2698.

Kalatian, A., & Shafahi, Y. (2016). Travel mode detection exploiting cellular network data. In *MATEC Web of Conferences* (Vol. 81, p. 03008). EDP Sciences.

Kido, H., Yanagisawa, Y., & Satoh, T. (2005, July). An anonymous communication technique using dummies for location-based services. In *ICPS'05. Proceedings. International Conference on Pervasive Services, 2005.* (pp. 88-97). IEEE.

Krumm. J. (2009). A survey of computational location privacy. Personal Ubiquitous Computing 13, 6 (Aug. 2009), 391–399. DOI:http://dx.doi.org/10.1007/s00779-008-0212-5

Kujala, R., Aledavood, T., & Saramäki, J. (2016). Estimation and monitoring of city-to-city travel times using call detail records. *EPJ Data Science*, *5*(1), 6.

Langford, M., Higgs, G., Radcliffe, J., & White, S. (2008). Urban population distribution models and service accessibility estimation. *Computers, Environment and Urban Systems*, *32*(1), 66-80.

Levinson, D. M., & Kumar, A. (1994). The rational locator: why travel times have remained stable. *Journal of the american planning association*, *60*(3), 319-332.

Little, R. J. (1993). Post-stratification: a modeler's perspective. *Journal of the American Statistical Association*, *88*(423), 1001-1012.

Louail, T., Lenormand, M., Ros, O. G. C., Picornell, M., Herranz, R., Frias-Martinez, E., ... & Barthelemy, M. (2014). From mobile phone data to the spatial structure of cities. *Scientific reports*, *4*, 5276.

Maantay, J.A., Maroko, A.R., and Herrmann, C., (2007). Mapping population distribution in the urban environment: the Cadastral-based Expert Dasymetric System (CEDS). Cartography and Geographic Information Science, 34 (2), 77–102. doi:10.1559/152304007781002190

Mennis, J. and Hultgren, T., (2006). Intelligent dasymetric mapping and its application to areal interpolation. Cartography and Geographic Information Science, 33 (3), 179–194. doi:10.1559/152304006779077309

Mir, D. J., Isaacman, S., Cáceres, R., Martonosi, M., & Wright, R. N. (2013, October). Dp-where: Differentially private modeling of human mobility. In *2013 IEEE international conference on big data* (pp. 580-588). IEEE.

Ortúzar S., J. d. D. & Willumsen, L. G. (2011). *Modelling transport* (4th ed.). Oxford: Wiley-Blackwell.

Phithakkitnukoon, S., Horanont, T., Di Lorenzo, G., Shibasaki, R., and Ratti, C. (2010). Activity -Aware Map: Identifying human daily activity pattern using mobile phone data, Human Behavior Understanding, 6219(3), 14-Springer Berlin / Heidelberg.

Phithakkitnukoon, S., and Ratti, C., (2011), Inferring Asymmetry of Inhabitant Flow using Call Detail Records, Journal of Advances in Information Technology, 2(4), 239-249.

Phithakkitnukoon, S., Smoreda, Z., and Olivier, P., (2012). Socio-geography of human mobility: a study using longitudinal mobile phone data. PLoS ONE, 7 (6), e39253. doi:10.1371/journal.pone.0039253

Qu Y, Gong H, Wang P (2015). Transportation mode split with mobile phone data. In: Intelligent transportation systems (ITSC), 2015 IEEE 18th international conference on. IEEE Press, New York, pp 285–289

Reades, J., Calabrese, F., and Ratti, C. (2009). Eigenplaces: analyzing cities using the space-time structure of the mobile phone network, Environment and Planning B: Planning and Design, 36(5), pp. 824-836.

Ruther, M., Leyk, S., and Buttenfield, B.P., (2015). Comparing the effects of an NLCD-derived dasymetric refinement on estimation accuracies for multiple areal interpolation methods. Giscience & Remote Sensing, 52 (2), 158–178. doi:10.1080/15481603.2015.1018856

Schneider, C. M., Belik, V., Couronné, T., Smoreda, Z., & González, M. C. (2013). Unravelling daily human mobility motifs. *Journal of The Royal Society Interface*, *10*(84), 20130246.

Sevtsuk, A., and Ratti, C. (2010). Does Urban Mobility Have a Daily Routine? Learning from Aggregate Data of Mobile Networks, Journal of Urban Technology, 17(1), 41-60.

Shad, S. A., & Chen, E. (2012). Precise location acquisition of mobility data using cell-id. *arXiv preprint arXiv:1206.6099*.

Silm, S. and Ahas, R., (2010). The seasonal variability of population in Estonian municipalities. Environment and Planning A, 42 (10), 2527–2546. doi:10.1068/a43139

Simini, F., Gonza´lez, M. C., Maritan, A., and Baraba´si, A. L. (2012). A universal model for mobility and migration patterns, Nature, 484, 96–100.

Smith, A., Martin, D., and Cockings, S., (2014). Spatio-temporal population modelling for enhanced assessment of urban exposure to flood risk. Applied Spatial Analysis and Policy, 9 (2), 145–163. doi:10.1007/s12061-014-9110-6

Smoreda, Z., Olteanu-Raimond, A. M., & Couronné, T. (2013). Spatiotemporal data from mobile phones for personal mobility assessment. *Transport survey methods: best practice for decision making*, *41*, 745-767.

Song, C, Koren, T, Wang, P, and Barabási, A. L. (2010a). Modelling the scaling properties of human mobility, Nature Physics, 6, 818–823.

Song, C., Qu, Z., Blumm, N., & Barabási, A. L. (2010b). Limits of predictability in human mobility. *Science*, *327*(5968), 1018-1021.

Stewart, D. W., & Shamdasani, P. N. (2014). *Focus groups: Theory and practice* (Vol. 20). Sage publications.

Steenbruggen, J., Borzacchiello, M. T., Nijkamp, P., & Scholten, H. (2013). Mobile phone data from GSM networks for traffic parameter and urban spatial pattern assessment: a review of applications and opportunities. *GeoJournal*, *78*(2), 223-243.

Traag, V. A., Browet, A., Calabrese, F., & Morlot, F. (2011, October). Social event detection in massive mobile phone data using probabilistic location inference. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing* (pp. 625-628). IEEE.

Traficom (2017). Telecommunications Markets in the Nordic and Baltic Countries. https://www.traficom.fi/sites/default/files/media/file/Telecommunications_Markets_in_the_Nordic_and_Baltic_Countries_2017.pdf

Traficom (2018). Turnover of Telecommunications Operations. Retrieved from: https://www.traficom.fi/en/statistics/turnover-telecommunications-operations

Trasarti, R., Olteanu-Raimond, A. M., Nanni, M., Couronné, T., Furletti, B., Giannotti, F., ... & Ziemlicki, C. (2015). Discovering urban and country dynamics from mobile phone data with spatial correlation patterns. *Telecommunications Policy*, *39*(3-4), 347-362.

Tolouei, R., Álvarez, P., & Duduta, N. (2015, September). Developing and verifying origin-destination matrices using mobile phone data: the llitm case. In European Transport Conference (Vol. 2015).

Tolouei, R., Psarras, S., & Prince, R. (2017). Origin-destination trip matrix development: Conventional methods versus mobile phone data. *Transportation research procedia*, *26*, 39-52.

United Nations (2019). Population Density. Retrieved from: https://population.un.org/wpp/Download/Standard/Population/

U.S. Department of Transportation, Federal Highway Administration (2017). National Household Travel Survey. Retrieved from https://nhts.ornl.gov.

VR (2019). Kaukoliikenteen reittikartta (in Finnish only). Retrieved from: https://www.vr.fi/cs/vr/fi/kaukoliikenteen-reittikartta

Väylä (2019). LAM – liikenteen automaattinen mittausjärjestelmä. Retrieved from: https:// osoite

Wang, P., Hunter, T., Bayen, A. M., Schechtner, K., and González, M. C. (2012). Understanding Road Usage Patterns in Urban Areas. Scientific reports, 2.

Wang, Z., He, S. Y., & Leung, Y. (2018). Applying mobile phone data to travel behaviour research: A literature review. *Travel Behaviour and Society*, *11*, 141-155.

Wesolowski, A., Eagle, N., Tatem, A. J., Smith, D. L., Noor, A. M., Snow, R. W., & Buckee, C. O. (2012). Quantifying the impact of human mobility on malaria. *Science*, *338*(6104), 267-270.

Wesolowski, A., O'Meara, W. P., Tatem, A. J., Ndege, S., Eagle, N., & Buckee, C. O. (2015). Quantifying the impact of accessibility on preventive healthcare in sub-Saharan Africa using mobile phone data. *Epidemiology (Cambridge, Mass.)*, *26*(2), 223.

White J, Wells I (2002). Extracting Origin Destination Information from Mobile Phone Data, Proc. IEEE RTIC, London, UK

Widhalm, P., Yang, Y., Ulm, M., Athavale, S., & González, M. C. (2015). Discovering urban activity patterns in cell phone data. *Transportation*, *42*(4), 597-623.

Williams, Nathalie E., Timothy A. Thomas, Matthew Dunbar, Nathan Eagle, and Adrian Dobra. (2015). "Measures of Human Mobility Using Mobile Phone Records Enhanced with GIS Data." PLoSONE 10 (7): e0133630. doi:10.1371/journal.pone.0133630. http://dx.doi.org/10.1371/journal.pone.0133630.

Wolf J, Oliveira M, Thompson M (2003). Impact of underreporting on mileage and travel time estimate – results from Global Postionning System enhanced household survey, *Transportation research record*, 1854, pp. 189-198.

Wu, L., Yang, B., & Jing, P. (2016). Travel mode detection based on gps raw data collected by smartphones: a systematic review of the existing methodologies. *Information*, *7*(4), 67.

Wu, S., Qiu, X., and Wang, L., (2005). Population estimation methods in GIS and remote sensing: a review. GIScience & Remote Sensing, 42 (1), 80–96. doi:10.2747/1548-1603.42.1.8

Zagatti, G. A., Gonzalez, M., Avner, P., Lozano-Gracia, N., Brooks, C. J., Albert, M., ... & Tatem, A. J. (2018). A trip to work: Estimation of origin and destination of commuting patterns in the main metropolitan regions of Haiti using CDR. *Development Engineering*, *3*, 133-165.

Zang, H., Baccelli, F., & Bolot, J. (2010, March). Bayesian inference for localization in cellular networks. In *2010 Proceedings IEEE INFOCOM* (pp. 1-9). IEEE.

Zhang, Y., Qin, X., Dong, S., & Ran, B. (2010, January). Daily OD matrix estimation using cellular probe data. In 89th Annual Meeting Transportation Research Board (Vol. 9).

Xu, B., Sun, G., Yu, R., & Yang, Z. (2012). High-accuracy TDOA-based localization without time synchronization. *IEEE Transactions on Parallel and Distributed Systems*, *24*(8), 1567-1576.

Xu, D., Song, G., Gao, P., Cao, R., Nie, X., & Xie, K. (2011, December). Transportation modes identification from mobile phone data using probabilistic models. In *International Conference on Advanced Data Mining and Applications* (pp. 359-371). Springer, Berlin, Heidelberg.