Enriching public transportation data using Bayesian methods

Philip Lemaitre

Advisor: Jes Frellsen
Co-Advisor: Michael Riis Andersen
Submitted: March 2022

IT University
of Copenhagen

Abstract

Public transportation has gone from data-poor to data-rich through the widespread use of Automatic Data Collection (ADC) systems such as the Automatic Fare Collecting (AFC), Automatic Vehicle Location data (AVL) and Automatic Passenger Count (APC) systems. These systems are designed for specific purposes: the AVL system monitors the public transit agencies' fleet, the AFC system collects revenue with financial accountability in mind, and the APC system counts the number of passengers in the vehicle at the stop level. These systems have separately and together contributed to a flourishing of valuable insights in public transportation research.

Nevertheless, some information remains unavailable to transit agencies due to the systems not collecting the information of interest, or agencies lacking access to the relevant systems. In the case of limited information in the data, transit agencies face the challenge of utilising the full potential of the information from these systems. This thesis focuses on the data from the AFC system generated with smart cards and how a Bayesian framework can infer the missing information of interest. The Bayesian framework has been found useful for handling this challenge since it makes it possible to model the complete data-generating process, even with sparse data, and even when it is possible to observe only parts of the data-generating process. The use of the Bayesian framework is demonstrated in the thesis by two published papers and a exploratory study.

The first paper investigates the case of not being able to access the recorded timetable information from the AVL system, and how using scheduled timetable information can affect train-to-passenger assignments. The paper presents a hierarchical Bayesian mixture model to infer the latent arrival times.

The second paper focusses on the challenge of the information of interest not being stored by the system. In this case, that is the activity of travellers transferring from bus to trains. When this information is not available, it is difficult for transit agencies to evaluate whether scheduled transfer times between vehicles are reasonable, since travellers could have engaged in some activity such as shopping, buying coffee, etc., affecting the observed distribution of walking times. The paper proposes a hierarchical Bayesian mixture model to infer latent behaviour, making it possible to infer the walking time distributions of walking directly and conducting an activity during the transfer.

Finally, this thesis contains an exploratory study. The study investigates the possibility of combining smart card data with journey planner search data to identify areas of interest, these being areas where people want to go, but which are not supplied by public transportation.

This PhD thesis presents new methods for using a Bayesian framework to infer missing data whose absence originates in a-priori system design.

Resume (Danish)

Offentlig transport har bevæget sig fra at være et data fattigt, til et data rigt område gennem den udbredte brug af Automatic Data Collection-systemer (ADC) som fx Automatic Fare Collecting system (AFC), Automatic Vehicle Location data (AVL) og Automatic Passenger Count Systems (APC). Disse systemer er designet til specifikke formål, med AVL-systemet designet til at overvåge de offentlige transportselskabers køretøjsenheder. AFC systemet indsamler indtægter med økonomisk ansvarlighed som primært formål, og APC-systemet tæller antallet af passagerer i køretøjet ved stoppestedet. Disse systemer har hver for sig og sammen, bidraget til en opblomstring af værdifuld indsigt indenfor forskningsfeltet.

Ikke desto mindre er nogle oplysninger stadig utilgængelige for transportselskaberne på grund af designet af systemerne, som ikke indsamler den relevante information eller pga. manglende adgang til de relevante systemer. I tilfælde af begrænset eller manglende information i data, står transportselskaberne overfor udfordringer ift. at udnytte det fulde potentiale fra disse systemer. Denne afhandling fokuserer på data fra AFC-systemet genereret med rejsekort data og hvordan en Bayesiansk ramme kan udlede den manglende information. Den Bayesianske metode har vist sig at være nyttig til håndtering denne udfordring, da den gør det muligt at modellere den komplette data generende proces, og virker selv med sparsomme data, samt i tilfælde hvor det kun er muligt at observere dele af den data generative proces. Brugen af den Bayesianske metode demonstreres i afhandlingen i form af to publicerede artikler og et studie.

Den første artikel undersøger tilfælde hvor det ikke er muligt at tilgå de registrerede køreplansoplysninger fra AVL-systemet og hvordan brugen af køreplaner kan påvirke tog-til-passager allokering. Artiklen præsenterer en hierarkisk Bayesiansk mixture model til at udlede de latente ankomsttider.

Den anden artikel fokuserer på udfordringen ved oplysninger om skiftertider, der ikke er lagret af systemet. Heriblandt aktiviteten af rejsende, der skifter fra bus til tog. Når denne oplysninger ikke er tilgængelige, er det vanskeligt for transportselskaberne at vurdere om de planlagte skiftertider mellem køretøjer er realistiske. Dette er som følge af at rejsende kan have lavet en aktivitet mellem de to afgange såsom shopping, køb af kaffe osv., hvilket har betydning for deres observede skiftetider. Artiklen foreslår en hierarkisk Bayesiansk mixture model til at udlede latent adfærd, og gør det muligt at udlede den direkte gangtid og gangtiden hvis den rejsende laver en aktivitet mellem afgangene.

Til sidst indeholder ph.d.-afhandlingen et udforskende studie. Studiet undersøger muligheden for at kombinere rejsekort data med rejseplanens søgedata for at identificere interesseområder, der defineres som områder, hvor folk gerne vil hen til eller fra, men er ikke understøttet af offentlig transport. På baggrund af de publicerede artikler har denne ph.d.-afhandling præsenteret nye metoder, hvori en Bayesiansk ramme er blevet brugt til at udlede den manglende data, der stammer fra design udfordringer i AFC-systemet.

Contributions

Paper A P. Lemaitre, M. R. Andersen and J. Frellsen, “When Did the Train Arrive? A Bayesian Approach to Enrich Timetable Information Using Smart Card Data”, has been published in IEEE Open Journal of Intelligent Transportation Systems, vol. 2, pp. 160–172, 2021, doi: 10.1109/OJITS.2021.3094620.

Paper B M. Eltved*, P. Lemaitre*, and N. C. Petersen*, “Estimation of transfer walking time distribution in multimodal public transport systems based on smart card data”, has been published in Transportation Research Part C: Emerging Technologies, vol. 132 , 2021 doi: 10.1016/j.trc.2021.103332.

Study P. Lemaitre, M. R. Andersen and J. Frellsen, Identifying areas of interest, exploratory study.

*Equal contribution by authors, see author statements.

Acknowledgements

First, I would like to thank the reader for taking the time to read my thesis. Completing this project has only been possible due to the help and support of others, which I am eternally thankful for.

Thank you to my supervisor Jes Frellsen for planting the idea of a PhD and guiding me into the Church of Bayes; I am now a true believer. Also, a great thanks to Michael Riis Andersen who, throughout my research stay at DTU, has been my unofficial co-supervisor, and has given me the crucial knowledge of how to implement models efficiently and correctly in STAN. Without the funding from the Innovation fund and Rejsekort & Rejseplanen A/S, the project would never have started; for that, I am very grateful. Thank you to my colleagues at Rejsekort & Rejseplanen for their support, especially to my company supervisor Signe Asser Møller for reminding me to take vacations, even though I have not been doing very well at following that advice. And thank you to Fares Bouanik for the crucial help with coming up with a great pitch to present the PhD to Rejsekort & Rejseplanen A/S and for the donation of a Mocca Master, which has provided necessary and essential coffee to me during my PhD—especially during the Covid shutdown.

To my friends Mikkel, Lasse, Barbara, Christian, Jep, Søren, Morten, Kasper and Hadlur, I owe a great deal of gratitude for your patience with me and for politely faking your interest in my thesis?Also, thank you to my PhD therapy group Mille and Mathias for providing a shoulder to cry on, sometimes this taking place through the screen during lockdown periods. Thank you to Else and Jens Peder for their effort of reading and trying to understand my articles, and to Irene, Simon, Jo (Johanne) and Solveig for their support and positive vibes.

To my parents, Lene and Pierre, I owe much gratitude for your everlasting support and belief in my ability to finish the project. Thank you to my sister-in-law Stephanie and my brother Nicklas for their encouragement and for proofreading the PhD, especially Stephanie with her sharp eyes and great insight. And to my nephews Arthur and Valdemar for their emergency care packet, which arrived at an essential time.

And last but not least, thank you to my lovely Asta. I could not have finished the thesis without her loving encouragement and support, for which I can not describe in words.

Contents

Chapter 1

Introduction

For decades, tickets in the public transportation sector have been issued using physical tokens, and transit agencies and transportation research have relied on manually collected data to understand public transportation use (Welch and Widita, 2019). This has changed with the emergence and widespread use of smart cards and Automatic Fare Collecting systems (AFC). With over 350 smart card systems around the world (Kurauchi and Schmöcker, 2017), the data that these systems generate have allowed researchers to analyse the public transportation sector in greater detail. As an example, researchers have used AFC data to measure the difference between passengers' scheduled travel time and actual travel time (Zhao et al., 2013), to measure how resilient a metro network is to disruptions (Jin et al., 2014), and to characterise station usage by passengers' travel habits (El Mahrsi et al., 2017).

The uses of AFC data are extensive. Nevertheless, there is still a need for more knowledge on the limitations and quality of data from the AFC system, with a common challenge being information missing from the data (Kurauchi and Schmöcker, 2017). Information can be missing when it is not collected, or through loss or corruption.

Robinson et al. (2014) have documented causes of lost or corrupted data in the AFC system. These causes can come from behavioural aspects, where travellers tap out too early or forget to tap out during the trip. Losses can also occur due to hardware issues, such as a broken card-reader or other device in the system, leaving travellers unable to tap in or out. These types of missing information are not systematic because they do not affect all the data collected by the AFC system. However, there is also systematically missing information in AFC data, originating in the design of the system.

AFC systems are designed to monitor and collect revenue, which means that the gathered data must be preprocessed before analysis for other purposes (Pelletier et al., 2011; Iliopoulou and Kepaptsoglou, 2019). A well known challenge due to the AFC design is destinations missing from trips Li et al. (2018). When travellers have to tap in only when entering the transportation mode, the trip's destination is missing. The missing destination creates a challenge if a transit agency is interested in building an origin–destination (OD) matrix to understand travel demand (Barry et al., 2009). In some AFC systems, the traveller has to tap in when changing transportation mode, but not when transferring between the same transportation mode, leading to missing observations (He and Trépanier, 2015). In cases where transfers are part of the trip, we may not observe other activities done by the traveller, like buying coffee or a newspaper, which may make the transfer time seem longer.

The design challenge results in a need to develop methods that can infer the missing information since the design affects the data as a whole, and therefore cannot be addressed by simply removing the data. The Bayesian framework is chosen as method for tackling this challenge here since Bayesian methods are able to encode domain knowledge into the model, which is suitable when data is sparse and has missing information.

1.1 Aim, objectives and research questions

The PhD thesis aims to explore missing information originating from the design challenges in AFC data, and to investigate how new methods developed using a Bayesian framework can infer the missing information.

To fulfil these aims, the thesis has two objectives:

1. To define and explore the missing information from the design challenge. The thesis achieves this by building on the AFC system challenges identified by Robinson et al. (2014), which is explored in section 1.3, with subsection 1.3.2 focussing on the Danish AFC system.

2. To develop new methods using a Bayesian approach for inferring the missing information originating from the design challenge. Chapters 2, 3 and 4 each propose a new method using a Bayesian approach to infer missing information.

The thesis contain three studies with the following research questions:

- Paper A: How do missing recorded arrival times affect the passenger to train assignment, and how can a Bayesian approach infer the missing recorded arrival times of trains?

- Paper B : How can the direct walking time during a bus-to-train transfer be inferred using a Bayesian approach, and how can the method be validated without a ground-truth data set?

- Exploratory study: Can the *areas of interest*—areas that people want to travel to or from but are not served by public transportation—be inferred using smart card data (AFC) and online search data?

1.2 Overview of the thesis

The thesis is divided into five chapters, with the rest of this chapter 1 containing three sections: in section 1.3, the missing information originating from the design of the system is explored, with section 1.4 introducing the Bayesian framework used in this thesis and section 1.5 summarising the published papers and the exploratory study of this PhD thesis.

In the chapters 2 and 3, respectively, the published Paper A and Paper B are presented. In chapter 4, an exploratory study on combining smart card data with journey planner data for the purpose of identifying the areas of interest is presented, and, lastly, in chapter 5, the conclusion of the thesis is presented.

1.3 Missing information in the AFC system—a view from the data generation process.

The smart card data from the AFC system provides the possibility of analysing the use of public transportation. Nevertheless, the information generated is not perfect. As with all real-world data, it contains errors and is not complete, as the information is limited by the system's design and by how the system is used by the end user (the traveller). This creates difficulties in obtaining data that accurately and sufficiently describes the public transportation system as a whole. These difficulties can originate in a range of challenges. Robinson et al. (2014) categorised the challenges of AFC systems data into four categories: *Hardware*, *Software*, *Data* and *User*. In this PhD thesis, matters associated with the business rule and the design of the system, which Robinson et al. (2014) categorises under *Software*, have been given their own category—*Design*, and the Robinson's *Data* is renamed *Input*. This provides the following categories, each of which is associated with different challenges:

- *Design* challenges originate in how the system is designed and the business rules defining how the traveller should use the system. The system does not collect information on the purpose of the trip, since the AFC is not designed to collect it. An example of a design challenge are the rules for tapping in and out. If the traveller has to tap in only when boarding the vehicle, the system will not collect information on when the traveller alights.

- *Software* challenges arise from the software, such as bugs, hacking of the systems, or software limitations. A possible limitation is the processing speed of the software, which can make it unfeasible to provide real-time information like passenger loads on the vehicles in current operation.

- *Hardware* challenges relate to the hardware used in the system. An example is broken card readers or smart cards, making it impossible for a traveller to tap in or out during part of the trip.
- *Input* challenges emerge from the input data to the system. This may include any form of data erroneously inputted to the system, like misspelt names of stops, or wrong timetable information. Vehicle routes are continually changed during the year, and, if the AFC system is not given correct information about the changes, then trips using a new route can be assigned incorrectly to a different route.
- *User* challenges arise from the traveller′s behaviour, such as a traveller forgetting to tap out, whether deliberately or accidentally, or a passenger using multiple cards or sharing a card. When a traveller has multiple cards or shares the card with others, the observed travel pattern over time will be ambiguous.

When encountering *Hardware*, *User* or *Input* challenges, a common approach is to exclude problematic data from the analysis (Li et al., 2018; Barry et al., 2009; Nassir et al., 2011; Alsger et al., 2016; Sánchez-Martínez, 2017; Dixit et al., 2019), since such data accounts for only a small fraction of the total. However, when it comes to the *Design* challenges of the AFC system, systematic limitations affect the system as a whole.

1.3.1 The design challenges

An AFC system's *design* determines the amount and type of information that it stores. The amount of data is determined by the number of interactions between a smart card and the AFC systems card readers, since each interaction will create a record (for further details, see Pelletier et al. 2011). Each record obtains its information from the smart card and the card reader, meaning that the stored data can only contain information generated from these sources. Other activities affecting the trip, such as buying coffee or shopping during a trip, which can affect the transfer time, are not captured by the AFC system. As a consequence, the collected information is limited to information on travellers trips inside the public transportation network. In the following subsections, the various aspects of the design challenges of AFC system are presented.

Outside and inside the public transportation network

The data collected by the AFC system is restricted to inside the public transportation network, which can be illustrated by the trip of going from home to work in fig. 1.1. The AFC system observes the movements between points A and D since it is inside the system, where the

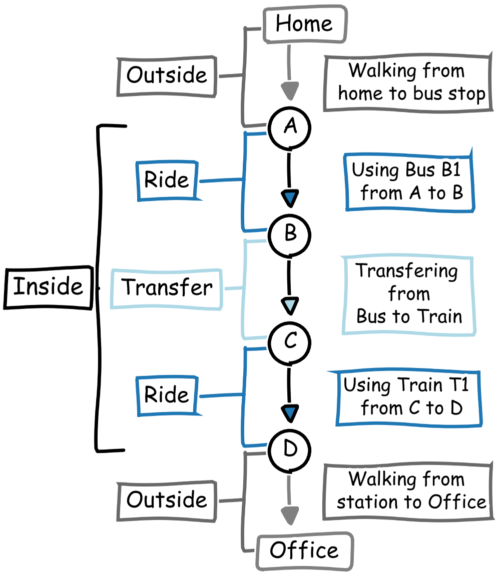

Home
Outside
Walking from home to bus stop
A
Ride
Using Bus B1 from A to B
B
Inside
Transfer
Transfering from Bus to Train
C
Ride
Using Train T1 from C to D
D
Outside
Walking from station to Office
Office

Figure 1.1: An illustration of how a trip from home to the office can be split into parts outside and inside the transit network. The part inside the transit network can be decomposed into rides and transfers. The trip describes the movement from A to D, where the movements from A to B and C to D are both rides since the traveller is using vehicles. The part from B to C is a transfer since the traveller changes vehicles (terminology from Robinson et al. 2014).

traveller can interact with the system. The AFC system cannot observe the movement outside the transportation network from home to point A, and from point D to the office, since the traveller cannot interact with the system during these legs. Thus, information such as movements outside the network and the trip's purpose is not collected. To gain this information, external data sources, such as land use, are needed to enrich the AFC data and to infer the trip's purpose (Alsger et al., 2018; Lu et al., 2018).

Type of system and placement of card-readers

When entering and moving inside the public transportation network, there are two major aspects of the AFC system that control the number of interactions and the information stored by the system. The first is the system type (open or closed AFC system), which affects the number of interactions, and the second is the placement of card readers in the system, which affects the information stored.

Open and closed system

Open systems are characterised by the traveller only having to use their smart card once per trip leg, either when boarding or alighting a transit vehicle (Kumar et al., 2018). Open systems hence do not record all locations in the trip chain, which is a well-known problem. The problem is handled by using trip chaining algorithms that combine the known location of the traveller's trip with timetable information to infer the missing links in trip (Nassir et al., 2011; Alsger et al., 2016; Sánchez-Martínez, 2017). Closed AFC systems are characterised by the traveller having to tap both in and out (Kumar et al., 2018), making it possible to generate a complete trip chain (van Oort et al., 2015; Yap et al., 2017; Dixit et al., 2019).

Placement of card readers

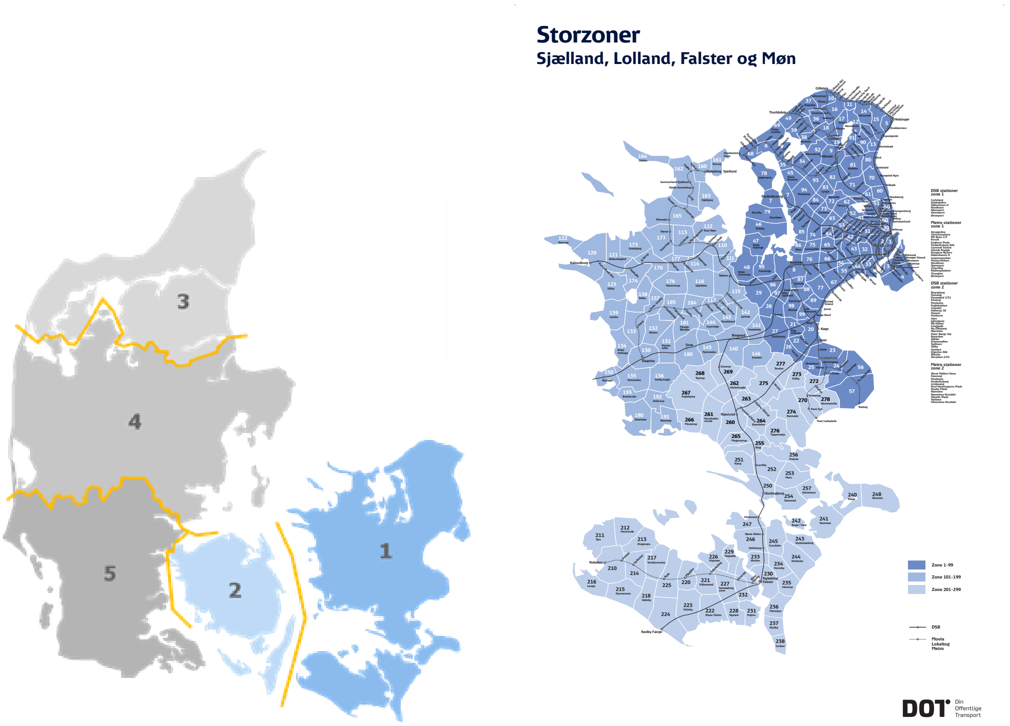
Card readers can be either location-placed or vehicle-placed. The placement affects the available information about the vehicle used and the waiting time before the a ride in a trip. Vehicle-placed card readers are located inside the vehicle, giving information on the specific vehicle and route used. Location-placed card readers are located at fixed locations. These can be at the platform or gates when entering or leaving the location. With location-paced readers, the vehicle id and route used will not be stored by the system. This makes difficult to associate trips with vehicles. Here, timetable information like AVL or General Transit Feed Specification (GTFS) can be used to enrich the smart card data (Luo et al., 2018). When it comes to the waiting time before the first ride of a trip, a vehicle-placed card-reader makes it impracticable to estimate the waiting time. With this placement, the traveller cannot tap in before the vehicle has arrived. The opposite is true for location-placed card readers since the traveller can tap in before the vehicle arrives, which makes it possible to study waiting time (Ingvardson et al., 2018).

1.3.2 The Danish AFC system and some of its design challenges

The Danish public transportation network is divided into regional zones as shown in fig. 1.2a), where each regional zone contains the local zone illustrated by fig. 1.2b. When a traveller uses public transportation, the trip's price will depend on the regional zone, the number of zones the traveller passes through, and which type ticket is used.

Ticket types

In the Danish transportation system, there are three main types of tickets: tickets issued by the various public transit agencies, single-use tickets, and the Danish smart card *Rejsekort* issued by the company *Rejsekort & Rejseplan A/S*. The tickets issued by *Rejsekort & Rejseplan A/S* are valid on all modes of public transportation, which makes it simple to travel within and between the different regions of Denmark. Due the smart cards only being a subset of the possible tickets, smart card-associated



(a) Danish transit system regional zones (Rejsekort & Rejseplan A/S, 2021b).

(b) Danish transit system regional zone 1 containing the local zones (DOT, 2021).

Figure 1.2: The Danish transit system regional zones and the regional zone 1 showing the local zones of that region.

data contain only a subset of all trips in the Danish public transportation network.

The Danish AFC company

The company *Rejsekort & Rejseplan A/S* manages all transactions in the national wide AFC system, which serviced over 140 million trips in 2019 with a turnover of 4.24 billion DKK (Rejsekort & Rejseplan A/S, 2021a). For the single-use tickets, information is limited since the single-use tickets only interact with the AFC system at the moment of issue. The main information collected for single-use tickets is the price, location and time of issuance, and the number of valid zones.

Different types of smart cards

Information associated with a smart card and the rules for using them affects the information generated whenever a smart card interacts with the Danish AFC system. For the Danish cards, the information depends on the type of smart card, since some cards follow the concept of a closed system, while others are similar to the open AFC system concept. Below are some of the most common types of cards and which type of system they follow (a more in-depth description of the different smart card types is provided in appendix A.):

- Similar to the open system: Commute, School and Youth Card (in Danish: Pendler-, Skole- og ungdomskort).
- Follows a closed system: Personal, Anonymous, Flex, Business and Commute-Combo card (Personligt-, Anonymt-, Flex-, Erhvervs- and Kombikort).

Cards that are similar to the open system

For the cards that are similar to an open system, the traveller only has to tap in when entering buses or when activating the card. This *design* feature limits the available data to the parts of a trip where a bus is involved. Despite this, these specific cards have further information relating to the trip and its purpose. The cards are specific to a single person, meaning the same person generates the data. The payment for the cards is a monthly fee for a determined number of zones, where the traveller is allowed to travel unlimited. The traveller has to be a frequent traveller for the payment plan to be cheaper than the closed-system cards. Since the trips are limited to specific zones, it would be reasonable to assume that these travellers are commuters. The trip purpose is more apparent for the card types youth and school, where the zones are limited to the home address and the place of education. The traveller is still allowed to use the card for other purposes if they are inside the zones. However, the *design* of these smart cards gives more information about the trip purpose than closed system cards.

Cards that follows a closed system

When using the Danish cards, which follows the closed system, travellers have to tap in at the start of the trip, when changing public transportation mode, and when ending the trip. With these cards, a traveller can travel anywhere with public transportation and pay per trip except when using a Combo-Commute card, which combines two fixed-fee payment types for pre-determined zones and per-trip payment outside these zones. Trips generated by these cards contain the trip-legs' times and locations, making it possible to monitor the travellers through the transportation network. With personal and Commute-Combo cards, it is possible to track the travel behaviour of the traveller over time since the use of the card is tied to its owner. In comparison, Anonymous, Flex and Business cards are not tied to one owner, meaning that several different people can have used the card, creating a mixed record of behaviours and travel patterns over time.

The placement of card-readers in the Danish AFC system

The Danish AFC system involves 18.876 card readers (Rejsekort & Rejseplan A/S, 2021a) around the country. The card readers are vehicle-placed for buses, and location-placed for trains (including trams and metro) at the station platforms. When a traveller tags in onboard a bus, the vehicle ID is stored, which means the vehicle used is known. In contrast, the vehicle is not stored when the traveller taps in at the train station, so the trains used by travellers need to be estimated.

The transportation network

An essential part of understanding how the transportation network is used is the transportation network itself. This is needed for combining smart card trips with vehicles used in the trip (Luo et al., 2018), However, the Danish AFC system only contains the scheduled timetable and not the recorded timetables, which means that AVL data is needed to give an accurate description of the system or methods for inferring the recorded timetables when they are not available.

Summary of identified Danish design challenges

- A design challenge can be an activity that affects the trip, but is not captured by the AFC system. An example is buying coffee or shopping.
- Travellers' movements outside the public transportation network are not recorded by the AFC system. An example is walking from the vehicle to the office, so there is no record to indicate that the trip purpose's is going to work.
- Since there are multiple types of tickets and multiple types of smart cards, the data collected from the smart cards will be only a subset of all trips in the public transportation network.
- To the describe the transportation network, the recorded timetable information is need to combine the smart trips with routes. However, the record timetable information is not stored in the AFC system.
- The locations of card readers affects what information is stored, and thereby affects what can be inferred from the data. Location-placed card readers, for example, do not record vehicle id's, making passenger-to-vehicle assignment challenging.

1.4 The Bayesian framework

In the Bayesian modelling framework, probability is used to measure the uncertainty of unknown quantities in the model. When assembling the model, all parts must be made as probabilistic statements with all connections to the unknown quantities expressed as a probability distribution (Ghahramani, 2004; Bernardo, 2011; Gelman et al., 2013). This makes it possible to quantify the uncertainty for all of the model's relevant quantities, such as the parameter uncertainty.

The model is constructed by assuming that data X is a random variable governed by an unknown parameter θ of interest. Since we do not know the true value of that parameter, it is treated as a random variable with the prior probability distribution $P(\theta)$. The relationship between the data X and the parameter θ can, by the use of Bayes' rule, be expressed as the conditional probability

$$\overbrace{P(\theta|X)}^{\text{Posterior}} = \frac{\overbrace{P(X|\theta)}^{\text{Likelihood}}\ \overbrace{P(\theta)}^{\text{Prior}}}{\underbrace{P(X)}_{\text{Evidence}}}. \tag{1.1}$$

In the Bayesian framework, the above equation is called *Bayes' theorem* and describes how the distribution of the parameter θ should be changed in light of new information relating to θ, which is expressed in the form of the data X.

The posterior distribution $P(\theta|X)$ describes the distribution of the parameter θ after observing the data X, whereas the prior distribution $P(\theta)$ describes the knowledge we have about θ before observing any data. For each possible value of θ, the likelihood $P(X|\theta)$ expresses how probable the data X is for a specific value of θ. For a fixed set of data X, the resulting likelihood distribution shows how the probability of the data X varies as a function of the parameter θ. It is important to note that likelihood is a probability distribution with respect to X and not θ, and thus will not be a probability distribution when integrating over the possible values of θ. However, the likelihood distribution is not a probability distribution since it does not integrate into one. The evidence $P(X) = \int P(\theta)P(X|\theta)d\theta$ is the marginal probability of the data and is often analytically intractable. Due to this, the posterior distribution is often inferred by approximation methods such as Markov Chain Monte Carlo (MCMC) computation. Thus the evidence $P(X)$ is sometimes omitted from the equation since it does not depend on the parameter of interest θ (Bishop, 2006; Gelman et al., 2013) and will then be proportional to the likelihood and the prior distribution, i.e.

$$P(\theta|X) \propto P(X|\theta)P(\theta). \tag{1.2}$$

After observing the data X and inferring the parameters θ, the distribution of potential new observations can be obtained from the predictive distribution given by

$$P(\widetilde{X}|X) = \int P(\widetilde{X}|\theta)p(\theta|X)\, d\theta. \tag{1.3}$$

By using a probabilistic approach, such as the Bayesian approach, all types of uncertainty—structural, parametric and measurement—are built into the model (Ghahramani, 2015), which forces the models to have explicit assumptions regarding the data-generating process. By describing the data-generating process explicitly, the models have more intuitive structures, making them and their parameters more interpretable than models such as neural networks, which can be challenging to interpret directly from their specifications. In addition to this, the Bayesian methods can often be applied to notably smaller data sets compared to classic methods such as maximum likelihood (ML)

$$\widehat{\theta}_{ML} = \underset{\theta}{\operatorname{argmax}}\, P(X|\theta) \tag{1.4}$$

This is possible due to the use of prior distributions, where the Bayesian framework encodes knowledge about the unknown parameter without losing the asymptotic properties as the classic approaches (Gelman et al., 2013).

The prior distribution

The prior distribution, or simply 'the prior', expresses our knowledge about the unknown parameters before we have observed any data. It can be specified with different purposes in mind. Non-informative priors such as Jeffreys or uniform priors are often used when little is known about parameter values or in order to affect the posterior as little as possible (Gelman et al., 2017). At the opposite extreme, informative priors are used when there is knowledge about the reasonable values of the parameters, minimising the effect of outliers on the posterior distribution (Gelman and Hennig, 2017). In-between the non-informative and informative priors lie the weakly informative priors. These priors are used to regularise the posterior distribution, thus avoiding overfitting to data through the likelihood. In addition, the informative and weakly informative priors can give smoother and more stable inference of the posterior distribution (Gelman et al., 2017). Other purposes relate to the form of the distribution, such as when using conjugated priors. These priors are often used for the convenience of the posterior distribution having the same functional form as the prior, making analysis simpler. In the case of hierarchical models, the prior encodes structural information by controlling the degree of information that are shared between the parameters in the model (Gelman et al., 2017).

An example of a prior

A simple way to illustrate the encoding of knowledge in the prior and the ability of Bayesian frameworks to handle small amounts of data is the classic example of a coin toss (following Gelman et al. 2013; Bayes 1763). If we flip N coins with the outcomes $X = [x_1, \ldots, x_N]$ and denote $\theta \in [0, 1]$ to be the probability of getting heads $x = 1$ and $1 - \theta$ being the probability of getting tails $x = 0$. Assuming that outcomes will be independent of the outcome of the previous coin flip, the data can be modelled as being conditionally independent and identically distributed.

It can then be assumed that the likelihood follows a binomial distribution with a conjugated Beta distribution prior, which gives the posterior distribution

$$\begin{aligned} P(\theta,\alpha,\beta|X) &= \mathrm{Bin}(X|\theta)\,\mathrm{Beta}(\theta|\alpha,\beta) \qquad (1.5)\\ &\propto \overbrace{\theta^{\sum x}(1-\theta)^{N-\sum x}}^{\text{Likelihood}}\,\overbrace{\theta^{\alpha-1}(1-\theta)^{\beta-1}}^{\text{Prior}}, \qquad (1.6)\end{aligned}$$

which can be reduced and expressed as

$$P(\theta,\alpha,\beta|X) = \mathrm{Beta}(\theta|H,T), \qquad (1.7)$$

$$\text{where } H = (\alpha+\textstyle\sum x) \text{ and } T = (\beta+N-\textstyle\sum x)$$

In the equation above, four things are affecting the inference of θ: the number of observed heads $(\sum y)$, the number of observed tails $(N-\sum y)$ and the hyperparameters α and β relating to our prior knowledge of θ. As H increases relative to T, the probability of getting heads increases vice versa. This setup makes it possible to interpret α and β as pseudo-coin flips based on knowledge of similar problems since the increases in, e.g. H can either come from observing a higher number of heads or higher values of α. To simplify and explain the effect, we can use the maximum a-posteriori (MAP) given the point estimate

$$\widetilde{\theta}_{\mathrm{MAP}} = \underset{\theta}{\operatorname{argmax}}\, P(X|\theta)P(\theta) \qquad (1.8)$$

We may draw on Gelman and Nolan (2002) who argues that there are no biased coins. With this knowledge, we may set a strong prior by setting α and β to equally high numbers, meaning that we need to see a high number of coin flips before there will be notable change from the estimate of $\widetilde{\theta}_{MAP} = 0.50$. On the other hand, if we do not know much about the coin flip, we may choose an uninformative prior by setting $\alpha = \beta = 1$, corresponding to a uniform prior. The estimation $\widetilde{\theta}_{\mathrm{MAP}}$ will quickly be dominated by the data and the prior will a small influence on the estimated $\widetilde{\theta}_{MAP}$. However, when we take only a few observations, say only three coin flips, all heads, the prior will have an influence θ. If we use the classical method of maximum likelihood (ML), the expected value will be $\widehat{\theta}_{\text{ML}} = 1$, which is a bit extreme when we have seen only three coin flips. When we use a weakly informed prior such as $\alpha = 2$ and $\beta = 2$, the MAP estimate value will be $\widetilde{\theta}_{\text{MAP}} = 0.8$, which is higher than the expected value of a fair coin, but lower than the ML estimation $\widehat{\theta}_{\text{ML}} = 1$.

1.4.1 Hierarchical models

In the previous section, we saw that a Bayesian model can be useful to infer the unknown parameter θ_1 even when there are only a few observations. To learn more about the probability of getting a head θ_1, we can flip the coin again and again to increase the number of observations X. However, suppose we lost the coin before we could flip it. How should we learn more about θ_1? We could use coins minted from the factory, where the first coin was minted. It would be reasonable to assume that these coins will exhibit similar behaviour as the first coin. Imagine that 100 newly minted coins from the factory, were each flipped 1,000 times. If all the coin flips turned out as heads, then it would be probable that the probability of θ_1 will display similar behaviour. Or, if all the coins show different behaviours, then the new coin flips will give inconsiderable information about the probability of θ_1.

The hierarchical model

A hierarchical model can describe the degree of similarities between the coins, where the variation between and within the coins are modelled. The idea is that the observed data X can be divided into J groups $\boldsymbol{X} = [X_1, \ldots, X_J]$, where the data in group $j \in \{1, \ldots, J\}$ is governed by the parameter θ_j describing the jth group. The hierarchical part (Gelman et al., 2013) is that the collection of parameter $\boldsymbol{\theta}$ is described by a shared parameter ϕ. One way to conceptualise the parameter ϕ is as prior for the parameters $\boldsymbol{\theta}$. This becomes clearer when writing out the joint probability as

$$P(\boldsymbol{\theta}, \phi) = P(\phi)P(\boldsymbol{\theta}|\phi) \tag{1.9}$$

$$= P(\phi)\prod_{j=1}^{J} P(\theta_j|\phi), \tag{1.10}$$

and, with Bayes theorem, we can write the joint posterior distribution as

$$P(\boldsymbol{\theta}, \phi|\boldsymbol{X}) \propto P(\boldsymbol{X}|\boldsymbol{\theta}, \phi)P(\boldsymbol{\theta}, \phi) \tag{1.11}$$

$$= P(\phi)\prod_{j=1}^{J} P(X_j|\theta_j)P(\theta_j|\phi). \tag{1.12}$$

The first thing to notice is that the likelihood of the data is only conditioned on $\boldsymbol{\theta}$ since the data X_j is independent of ϕ when θ_j is given (Gelman et al., 2013). The second thing is that the distribution of the group-level parameters θ depends on the prior ϕ. Yet, an observation of X_1 will indirectly inform the parameters $\theta_{j\neq 1}$. Since X_1 informs parameter θ_1, which informs the shared parameter ϕ, it hence informs all other parameters $\theta_{j\neq 1}$.

An example of the hierarchical concept

To understand this concept, we can imagine an example of travellers

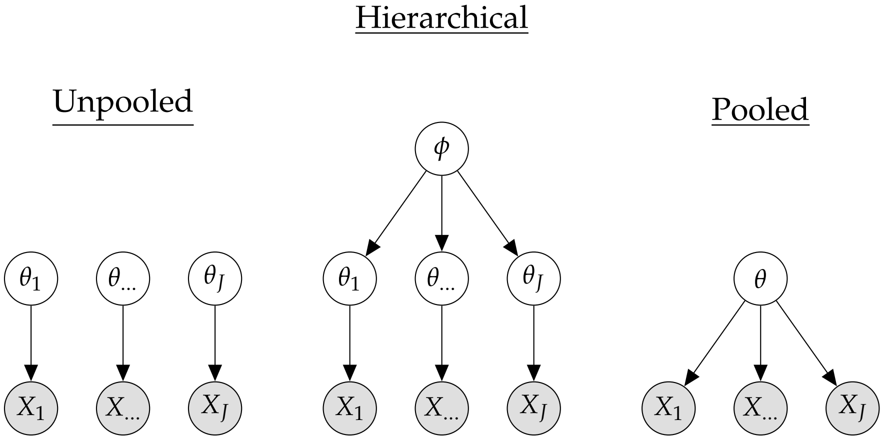


Figure 1.3: Examples of probabilistic graphical structural of the Unpooled, Pooled and Hierarchical model

transferring at a station with J different platforms, each with its own platform number $j \in (1, \ldots, J)$. For each platform, there will be a direct transfer time θ_j to a given point A, which will depend on the path from A to platform j. The transfer times $\boldsymbol{\theta}$ will vary somewhat, but there will be similarities due to the transfers being restricted by the station's layout, which will have a finite number of direct walking paths. All possible transfer times $\boldsymbol{\theta}$ between point A and the platform j will be expressed by the station-level characteristics ϕ.

At each platform, there will be N_j travellers with observed transfer times of X_j. Given that the station-level characteristics ϕ and the platform-level $\boldsymbol{\theta}$ transfer times cannot be observed, it is possible to model the data generation process with three different assumptions (Gelman et al., 2013). The corresponding approaches are described below and illustrated by fig. 1.3:

- *Pooled*: Assume that the transfer times $\boldsymbol{\theta}$ are identical by pooling the data together and modelling θ as a single distribution.

- *Unpooled*: Assume that the transfer times $\boldsymbol{\theta}$ are independent, and model each θ_j as a separate distribution using only the data from the jth platform to infer θ_j.

- *Hierarchical*: Assume that the $\boldsymbol{\theta}$ are conditionally independent given the station-level characteristics ϕ, and model each θ_j as a separate distribution conditional upon the distribution of ϕ.

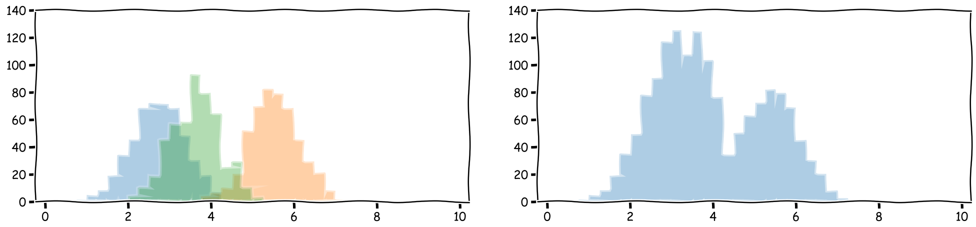
In the *pooled* model, the posterior of $\widetilde{\theta}$ will have to capture the variation between and within the transfer times of the different platforms. If we have two platforms $J = 2$ and there is a greater difference between the transfer times θ_1 and θ_2, then the posterior of $\widetilde{\theta}$ will be dominated by the θ_j with the greatest sample size N_j.

In the opposite *unpooled* model, the posterior of $\widetilde{\theta}$ will only capture the variation within each group, thereby not capturing the information between the groups. In addition to this, both the *unpooled* and *pooled* model cannot simulate new platforms of $\widetilde{\theta}$ since the models do not have information on the distribution of $P(\phi)$.

In between these two models lies the *hierarchical* model, where the station-level characteristic ϕ is inferred, making it possible to model the variation both between and within the groups. When the transfer times $\boldsymbol{\theta}$ between platforms are similar, meaning low variation between the platforms, the *hierarchical* model will mimic the *pooled* model. When the transfer time $\boldsymbol{\theta}$ between platforms are dissimilar, meaning high variation between the groups, the *hierarchical* model will mimic the *unpooled* model.

1.4.2 Mixture models

In the previous section, we could divide the data into different groups since there was data containing the information about the J groups. In such a case, data can easily be separated into distinct groups, as illustrated in fig. 1.4a. On the other hand, the information indicating group affiliations may not be available, making it difficult to separate the data into different groups, as illustrated in fig. 1.4b.



(a) Group information is available in the data.

(b) Group information is not available in the data.

Figure 1.4: Illustration of how the same data is distributed when the data have group information and when it does not.

If we know or assume that the N samples $X = [x_1, \ldots, x_N]$ originated from J different groups govern by distributions with parameters $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_J]$, then, for each sample $i \in \{1, \ldots, N\}$, there will be a corresponding unknown indication variable $Z = [z_1, \ldots, z_N]$ with $z_i \in \{1, \ldots, J\}$ indicating the group of the ith sample (Gelman et al., 2013; Ghahramani, 2013). Given this, the conditional distribution of x_i, given θ and z_i, can be expressed as

$$P(x_i|\theta, z_i = j) = P(x_i|\theta_{z_i=j}) = P(x_i|\theta_j). \tag{1.13}$$

Since all samples belong to a single group, we can express the proportion of samples in each group by $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_J]$ and the probability of ith sample belong to group j as

$$P(z_i = j|\boldsymbol{\lambda}) = \lambda_j, \tag{1.14}$$

$$\text{where} \sum_{j=1}^{J} \lambda_j = 1 \quad \text{and} \quad 0 \leq \lambda \leq 1.$$

With this, we can express the joint likelihood by marginalising over the indicator variable Z, such that

$$P(x_i|\boldsymbol{\theta}, \boldsymbol{\lambda}) = \sum_{j=1}^{J} \lambda_j P(x_i|\theta_j). \tag{1.15}$$

Using Bayes' theorem, we can express the posterior distribution as

$$\begin{aligned} P(\boldsymbol{\theta}, \boldsymbol{\lambda}|X) &\propto P(\boldsymbol{\theta}, \boldsymbol{\lambda})P(X|\boldsymbol{\theta}, \boldsymbol{\lambda}) & (1.16) \\ &= P(\boldsymbol{\theta}, \boldsymbol{\lambda}) \prod_{i=1}^{N} \sum_{j=1}^{J} \lambda_j P(x_i|\theta_j) & (1.17) \end{aligned}$$

where $P(\boldsymbol{\theta}, \boldsymbol{\lambda})$ is the joint prior, λ_j is the mixing component and $P(X|\theta_j)$ is the likelihood of the model. If we assume independence between $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$ then we obtain the graphical representation in fig. 1.5.

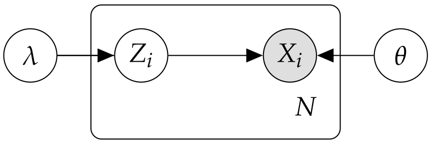


Figure 1.5: Graphical represented of a mixture model.

Challenge of using mixture models

The mixture models are a flexible class that make it possible to model complex structures. However, the flexibility can easily result in overfitting the data. A simple example is inferring a Gaussian mixture model with two components and N samples using the classical maximum likelihood method. When maximising the likelihood, a component can concentrate on a single data point. When this happens, the variance of this component will converge to zero leading the likelihood to go to infinity (Bishop, 2006). This behaviour is avoided in the Bayesian framework by the priors, which will regulate the variance the variance of the Gaussians.

Another challenge with the mixture models is the problem of identifiability (label switching), where the permutation of the J group labels of the model does not change the distribution Gelman et al. (2013) e.g. the likelihood is the same when changing the labels of θ_1 and θ_2.

$$\lambda_1 = 0.2 \quad \theta_1 = 2 \quad \text{and} \quad \lambda_2 = 0.8 \quad \theta_2 = 1 \tag{1.18}$$

is the same as

$$\lambda_1 = 0.8 \quad \theta_1 = 1 \quad \text{and} \quad \lambda_2 = 0.2 \quad \theta_2 = 2.$$

1.4.3 Approximation inference

Usually, we cannot obtain an analytic expression for the posterior distribution because the involved marginal distribution is a high-dimensional integral of a complex expression. In these cases, approximation methods are used, with the Markov Chain Monte Carlo (MCMC) method being the most widespread for Bayesian inference (Ghahramani, 2013).

The idea is that the posterior distribution can be simulated with the use of Markov chains, which are sequences of random variables $\theta^1, \theta^2, \ldots$ with the property that θ^t depends only on the previous θ^{t-1} through the conditional probability

$$P(\theta^t|\theta^{t-1}, \theta^{t-2}, \ldots) = P(\theta^t|\theta^{t-1}) \equiv T(\theta^t|\theta^{t-1}). \tag{1.19}$$

Here, $T(\theta^t|\theta^{t-1})$ is called the transition distribution. The Markov chain is designed in a way that the chain will converge asymptotically to a chosen target stationary distribution $P(\theta)$. A sufficient condition for the existence of a stationary distribution $P(\theta)$ is the detailed balance condition (Gelman et al. 2013; Bishop 2006)

$$P(\theta^t)T(\theta^{t-1}|\theta^t) = P(\theta^{t-1})T(\theta^t|\theta^{t-1}). \tag{1.20}$$

The following subsection describes MCMC methods, which satisfies the detailed balance.

Metropolis–Hastings (MH) algorithm

One of the more general MCMC algorithms is the Metropolis–Hastings (Metropolis et al., 1953; Hastings, 1970; Gelman et al., 2013), which produces a chain $\theta^0, \theta^1, \theta^2, \ldots$ by proposing a new state θ^* from a proposal distribution $Q(\theta^t|\theta^{t-1})$ and then, with some probability, accepting or rejecting the new state. The probability of accepting the state is constructed so that areas of the posterior distribution with higher densities get accepted more frequently than areas with lower density, thereby mimicking the posterior distribution. The pseudo-code for the Metropolis–Hastings algorithm is shown in algorithm 1. The beauty of the algorithm is its simplicity. However, the Metropolis–Hastings can experience difficulty exploring the posterior distribution in high dimensions. In addition to this, the step size determined by the proposal distribution $Q(\theta^t|\theta^{t-1})$ can be too small, resulting in inefficiencies that slow exploration of the posterior, whereas a too-large step size can lead to a high degree of rejections.

Hamiltonian Monte Carlo (HMC) and NUTS-sampler

The Hamiltonian Monte Carlo method is inspired by the movement of particles in physics, where the parameter of interest θ is seen as the particle's position, and an auxiliary variable m is introduced to describe the

Algorithm 1 Pseudo code for Metropolis–Hastings algorithm.

```
Initialise $\theta_0$
for $t = s$ to $S$ do
    1. Draw a proposal $\theta^*$ from a proposal distribution $Q(\theta^t|\theta^{t-1})$.
    2. Calculate the density ratio
        $$r = \frac{P(\theta^*|X)\ Q(\theta^{t-1}|\theta^*)}{P(\theta^{t-1}|X)\ Q(\theta^*|\theta^{t-1})}$$
    3. Draw value $\tau \in [0,1]$ from a uniform distribution.
    if $\tau < \min(1, r)$ then
        Accept proposal and set $\theta^t = \theta^*$.
    else
        Reject proposal and set $\theta^t = \theta^{t-1}$.
```

particle's momentum. By framing the setup in this way, it is possible to explore the target distribution $P(\theta|X)$ using Hamiltonian dynamics. To achieve this, we define the total energy of the system by the Hamiltonian function (Neal 2012; Betancourt 2017)

$$H(\theta, m) = U(\theta) + K(m) \tag{1.21}$$

with kinetic energy $K(m)$ and potential energy $U(\theta)$. The potential energy is conveniently defined as

$$U(\theta) = -log\,\overbrace{P(\theta|X)}^{\text{Posterior}} \tag{1.22}$$

and the kinetic energy is

$$K(m) = \frac{1}{2}m^T M^{-1} m \tag{1.23}$$

with M being a mass matrix (Gelman et al., 2013). The joint probability distribution of θ and m is then described by the Boltzmann distribution (Neal, 1994)

$$P(\theta, m) \propto e^{-H(\theta,m)}. \tag{1.24}$$

Combining the equations above, we can express the joint probability as

$$P(\theta, m) \propto e^{-H(\theta,m)}. \tag{1.25}$$

$$= e^{-U(\theta)-K(m)} \tag{1.26}$$

$$= P(\theta|X)e^{-\frac{1}{2}m^T M^{-1} m} \tag{1.27}$$

$$\propto P(\theta|X)N(m|0, M) \tag{1.28}$$

From this, we can explore the density of $P(\theta|X)$ by sampling from $N(m|0, M)$ and simulating/solving the Hamiltonian equations

$$\frac{d\theta}{dt} = \quad \frac{\partial H}{\partial m} = \frac{\partial K}{\partial m} \tag{1.29}$$

$$\frac{dm}{dt} = -\frac{\partial H}{\partial \theta} = \frac{\partial U}{\partial \theta} \tag{1.30}$$

describing how θ and m change over a fictitious time t. The Hamiltonian is simulated by numerical integration using the leapfrog algorithm. This algorithm discretises the dynamics of the equation into L steps of size ϵ Neal (2012). The pseudo-code for Hamiltonian Monte Carlo is described in algorithm 2. The downside of this implementation is that ϵ and L need to be specified. If ϵ is set too big, we will get a lot of rejections, and, if too small, the exploration will take a long time. If the number of steps L is too low, the exploration begins to imitate random-walk behaviour, whereas if it is too high, we risk making a U-turn in the exploration and ending up back where we started. The No-U-turn sampler (NUTS) solves this by dynamically setting the number of steps L and the step size ϵ for the leapfrog algorithm (Hoffman and Gelman, 2014).

Inference of mixture models

As mentioned before in section 1.4.2, mixture models can exhibit problems with identifiability (label switching), which can cause inference challenges for MCMC methods. The MCMC is an iterative method, and, due to the samples for the posterior distribution being obtained during this iterative process, the samples can undergo label switching. This challenge can be handled with different approaches such as enforcing

Algorithm 2 Pseudo-code for Hamiltonian Monte Carlo.

```
Initialise $\theta^0$, $\epsilon$, $S$, $L$
for $s = 1$ to $S$ do
    1. Draw $m^{t-1}$ from its posterior distribution $m \sim N(0, M)$.
    set $\theta^* = \theta^{t-1}$, $m^* = m^{t-1}$,
    for $l$ in $L$ do
        $\theta^*$, $m^* \leftarrow$ leapfrog($\theta^*$, $m^*$, $\epsilon$)
    3. Calculate the density ratio
            $$r = \frac{P(\theta^*|X)\ N(m^*|0, M)}{P(\theta^{t-1}|X)\ N(m^{t-1}|0, M)}$$
    4. Draw value $\tau \in [0, 1]$ from a uniform distribution.
    if $\tau < \min(1, r)$ then
        Accept proposal and set $\theta^t = \theta^*$.
    else
        Reject proposal and set $\theta^t = \theta^{t-1}$.
```

order on the relevant parameters—$\theta_1 < \theta_2 < \cdots < \theta_j$—or by permuting the samples in chains of the Markov chain Monte Carlo (MCMC), making the posterior distribution of each parameter unimodal (Stephens, 2000; Gelman et al., 2013; Papastamoulis, 2016).

1.4.4 Probabilistic programming

The idea of probabilistic programming is to combine the automation of programming with the language of statistical inference to represent probabilistic models. This can done by the probabilistic programming language having syntax for defining the probabilistic models and implementing the procedures for inferring the models automatically, instead of the user having to manually calculate or code the inference procedure (Ghahramani, 2015). The probability programming language Stan is designed for automatic inference in the Bayesian framework. Approximate inference methods such as MCMC are implemented, which means that the user only has to specify the models as a probability distribution.

1.5 Summary of studies

1.5.1 Paper A

"*When Did the Train Arrive? A Bayesian Approach to Enrich Timetable Information Using Smart Card Data*", studies how uncertainty about the arrival and departure times affects passenger-to-train assignments from smart card data, and proposes a novel method for inferring trains' arrival times when the actual timetable information is not available. If the recorded timetable information is missing and the scheduled timetable is used instead, the inaccuracy can induce errors in downstream analysis. The paper illustrates how the tap-in and tap-out distributions are altered and how passenger-to-train assignments are affected by using scheduled timetables (such as General Transit Feed Specification (GTFS)) rather than the actual timetable recorded in AVL data.

The paper proposes a hierarchical Bayesian mixture model to infer the missing arrival times using prior knowledge about how tap-outs (Hong et al., 2016) and train delays are distributed (Cerreto et al., 2018). Tap-outs are known to cluster together right after the arrival of a train (Hong et al. 2016; Min et al. 2016; Tan et al. 2021). However, the train used by any particular traveller is unknown. Thus, a mixture component is used to model the uncertainty of the train ridden by a traveller. The model can be used to identify the trains used by passengers and to classify the trains into early, on-time and late categories. The Bayesian model is evaluated on a Danish regional route with 15,136 trains and 51,933 trips, focussing on their arrivals at four stations. The method can infer 70% of the train arrivals with an average error of 30 to 42 seconds, depending on the station. The method is compared to state-of-the-art methods for inferring missing train arrivals using cross-validation. The paper contributes a method for inferring the missing recorded arrival times of trains and an examination of the consequences of using scheduled rather than actual times.

1.5.2 Paper B

"*Estimation of transfer walking time distribution in multimodal public transport systems based on smart card data*", investigates the direct walking-time distributions of travellers transferring from buses to trains. The direct walking time is relevant for transit agencies since it can be used to evaluate and improve connections between public transportation services (Parbo et al., 2014). The paper combines AFC and AVL data to infer the walking time of individuals transferring from a bus stop to a train platform. The observed walking times will be affected by individuals' walking paces (Daamen et al., 2006), the paths taken between the transfer locations (Daamen et al., 2006), and any activities that take place during a transfer (such as shopping, buying coffee etc.)

The paper proposes a hierarchical Bayesian mixture model to isolate and infer the walking times for travellers walking directly from a bus stop to the train platform. Due to the transfer activities of travellers not being observed by the AFC or AVL system, the model introduces a mixing component with two distributions: one to describe the walking time of travellers walking directly, and one to describe the walking time of travellers undertaking other activities during their transfers. At each station, the travellers transferring will be restricted by the same station layout, thus imposing some similarities between the walking paths at the station; this is modelled by a hierarchical element between the bus stops and the train platform.

The Bayesian model is assessed on 129 stations with a total of 1,145 combinations of bus stops and train platform validators. The inferred direct walking time distributions are compared to the scheduled transfer times, indicating sub-optimal connections between services at some stations. Due to the latent nature of the activities, the paper proposes two verification methods for evaluating the model results. The paper contributes a data-driven method for inferring the direct walking time distributions at scale.

1.5.3 Exploratory study

"*Chapter 4, Identifying areas of interest*" explores the possibility of combining smart card data (AFC) with journey planner search data to identify *areas of interest*. The study defines *areas of interest* as areas that people want to travel to or from but which are not served by public transportation. These places are of interest to public transit agencies since, if found, it may be possible to attract new travellers or increase the use of public transportation by existing users by enabling more people to go where they want to go.

As a quantification for *areas of interest*, the study proposes the use of ratios between online searches and the actual trips observed in the AFC system. The hypothesis is that higher ratios of searches to trips indicate possible *areas of interest*.

The hypothesis is investigated by using Danish municipalities as potential areas and inferring four ratios, two using the marginal probability and two using the conditional probability. For the ratios using the marginal probability, the first is the ratio between the probability of travelling *from* a municipality and the probability of searching for a trip *from* a municipality. The second ratio is between the probability of travelling *to* a municipality and the probability of searching for a trip *to* a municipality. The two other ratios are inferred by conditioning the probabilities respectively on starting *from* a municipality and going *to* a municipality. For inferring the probability, maximum likelihood is used for the marginal probabilities due to the simplicity of the probabilities involved. The conditional probabilities are inferred using a Bayesian model with multinomial distribution and a conjugated Dirichlet prior.

The study investigates the ratios using 98 Danish municipalities with 138 million smart card trips and 340 million online journey searches from 2018. The preliminary results show that there are municipalities with higher ratios, which could indicate potential *areas of interest*. However, the higher ratios may be due to other factors, such as behavioural factors tied to the difference in behaviours between searching and travelling.

With other factors that may explain the higher ratio, and without the ground truth for validating the inferred *areas of interest*, the proposed method of the study is inconclusive.

There is limited research on how smart card data and online journey planner data should be combined and for what purpose. This exploratory study contributes a proposal for identifying *areas of interest* from existing data and quantifying the levels of traveller interest in them, which can be used as starting point for further research.

1.5.4 Summary of contributions

In summary, the thesis contributes to the literature with the following:

- The thesis expands on the challenges of AFC data identified by Robinson et al. (2014) through the concept of the design challenge.
- The thesis has explored and showed how a Bayesian approach can be used to infer the missing information that originates from the design challenge with the following:
  - The thesis has examined how the use of scheduled timetable information instead of recorded timetable information can affect an analysis.
  - In cases of missing recorded timetable events, the thesis has contributed a Bayesian approach to infer the missing arrival times of trains.
  - The thesis has developed a data-driven and scalable Bayesian approach for inferring the direct walking-time distributions during passengers' transfers from buses to trains.
  - For situations when the ground truth for the transfer behaviour of travellers transferring from bus to trains is not available, the thesis has developed two new validation procedures for evaluating transfer behaviours.

- Explored the use of smart card data and online journey planner search data to identify the *areas of interest* with a proposed quantification and Bayesian approach for inferring the *areas of interest*.

Chapter 2

Paper A - When Did the Train Arrive? A Bayesian Approach to Enrich Timetable Information Using Smart Card Data

The following pages contain the article which has been published in the IEEE Open Journal of Intelligent Transportation Systems, vol. 2, pp. 160-172, 2021, doi: 10.1109/OJITS.2021.3094620.

IEEE Open Journal of
Intelligent Transportation
Systems

Received 26 February 2021; revised 18 May 2021; accepted 22 June 2021. Date of publication 5 July 2021; date of current version 27 July 2021.

Digital Object Identifier 10.1109/OJITS.2021.3094620

When Did the Train Arrive? A Bayesian Approach to Enrich Timetable Information Using Smart Card Data

PHILIP LEMAITRE[1], MICHAEL RIIS ANDERSEN[2], AND JES FRELLSEN[1,2]

[1]Department of Computer Science, IT University of Copenhagen, 2300 Copenhagen, Denmark

[2]Department of Applied Mathematics and Computer Science, Technical University of Denmark, 2800 Kongens Lyngby, Denmark

CORRESPONDING AUTHOR: P. LEMAITRE (e-mail: phle@itu.dk)

This work was supported by in part by the Danish Innovation Foundation under Grant 7038-00186B with co-funding from Rejsekort & Rejseplan A/S.

ABSTRACT Smart card data from the Automatic Fare Collecting systems (AFC) and timetable information, such as Automatic Vehicle Location (AVL), are used in combination by practitioners and researchers to gain a deeper understanding of the public transit network. In some cases, AVL data are not available due to records being missing in the system. In such cases, people resort to the used schedule timetable such as General Transit Feed Specification (GTFS) to match smart card data to the transit network. Since delays or changes to the timetable are not contained in the scheduled timetable, it can result in wrong matches between the smart card data and the transit network. This paper shows how the uncertainty of arrival and departure times affects passengers to train assignments and proposes a method for estimating the missing arrival time of trains when the recorded timetable information is not available. The method uses the knowledge of how the tap-outs are distributed in a hierarchical, latent Bayesian model to predict the arrival times of trains. Evaluated on 15,136 train arrivals, the model can infer 70% of the arrivals times with an average error of 28 to 32 seconds depending on the station.

INDEX TERMS AFC, automatic fare collection, AVL, automatic vehicle location, Bayes statistics, machine learning, missing data, smart card, train logs.

I. INTRODUCTION

THE COMBINATION of timetable information and Automatic Fare Collecting systems (AFC) is being used more broadly by researchers and practitioners to understand public transportation on the strategic, tactical and operational level [1], [2], [3]. This has lead to the identification of issues with the data used [1], [2], [3], [4], [5], [6], [7]. It has been shown that buses using AVL can have errors due to broken GPS units (hardware error), the bus deviating from the scheduled route (operational error) or busses not uploading data (data error) [5], which can lead to the travelers having the wrong alighting stop stored [4]. The same is apparent for trains, where AVL data can be missing for single trains or complete routes [6], [8], [9]. When recorded timetable information is not available, it is possible to use the scheduled timetable such as General Transit Feed Specification (GTFS) as a proxy. Using the scheduled timetable as a proxy leads to errors for the arrival and departure of the train, which can be due to the train being early, delayed or cancelled. The error, i.e., the difference between actual and scheduled arrivals of the trains, can propagate to erroneous estimates and conclusions in downstream analysis when using scheduled timetable information, such as the actual passenger trajectories [10], [11], [12], [13], the waiting time [14], [15] or the estimation of alighting stops using trip chaining [10], [16].

Different approaches have been taken to solve the issue of combining timetable information with smart card data. A method for fusing AVL, GTFS, and AFC data [6] for assigning passengers to trams in Hague solves the issues of scheduled timetables by two steps. First, passengers are assigned to each train ID using scheduled time with a bound allowing the tap to happen 20 to 50 seconds before or after the scheduled time, depending on it being a tap-in or -out. Then using the activity of the tap-in and -out to evaluate if the train is cancelled. However, by using a fixed bound, they exclude trains, which have larger delays.

The review of this article was arranged by Associate Editor Abdulla Hussein Al-Kaff.

In contrast, [9] and [17] estimate the complete missing timetable of large metros using non-parametric density estimation. Both methods can be divided into three main steps;

Step 1: Density estimation based on tap-outs.
Step 2: Train matching for the estimated densities from different stations.
Step 3: Estimation of departure and arrival time of trains.

In the first step, tap-outs are separated by a non-parametric density estimation using the knowledge that the tap-out pattern tends to cluster together in a clear dense pattern after the train arrivals. The densities are later in the third step used to infer the arrival time of the train. The first method [9] uses a histogram-based method called *S-Epoch* for estimating the density of the tap-outs to identify the points in time with the largest changes in the number of passengers aligning at a metro station. The second method [17] used *kernel density estimation* (DENCLUE 2.0) [18] to cluster the tap-outs. In the second step of both methods, a train matching algorithm is applied to connect the tap-out density estimation for the same train at different stations into a complete sequence. The third step is to infer the departure and arrival time of trains based on the density estimates [9], [17]. Both methods use the earliest tap-out from the densities to estimate the arrival time. However, for the departure, the *S-Epoch* assumes the trains depart immediately after arrival, where the *kernel density estimation* method uses the latest tap-in from each cluster of the train matching algorithm to infer the departure of the trains [17]. Finally, the inferred departure and arrival times are refined by shifting the estimations to account for the gate to train walking time.

Compared with previous approaches, our work focuses on the first step of density estimation and the third step of inferring the missing arrival times of trains by using a Bayesian framework. The Bayesian framework makes it possible to make a dynamic inference of the trains arrival time from the tap-out distributions instead of a static shift and links the scheduled timetable train IDs to tap-out distributions. The Bayesian model is evaluated on data from the Danish AFC system from February 1st to May 31st 2019, containing 51,933 trips and 15,136 intercity train arrivals for a regional route in Denmark. A more in-depth description of the data used is presented in Section IV-A. The data differ from the previous approaches [9], [17], which was applied to larger metros with gates, compared to the smaller Danish regional route without gates. The smallest average number of trips per trains was 7.0-8.0 [9], wherein our case study, the average trips per train range from 1.5-6.0 depending on the station.

Main contributions: Our main contributions are 1) to show how the use of erroneous timetable information can affect analysis and 2) how a Bayesian probabilistic framework can be used to infer missing arrival times.

Overview: The paper is structured in the following way; in the next Section II the problem of an erroneous timetable is illustrated, in Section III our proposed Bayesian model is derived and described, in IV a case study is conducted using the model on a Danish regional line and the final Section V contains the conclusion. To avoid any ambiguity, arrival and departures will only refer to the arrival and departure of trains, where alignment and boarding will refer to passengers.

II. THE EFFECT OF ERRONEOUS TIMETABLE INFORMATION

In this section, we will illustrate the consequence of using erroneous timetable information in an analysis. As a case study, we will use data from the Danish AFC system, which is an open system [16] with validation devices for tap-in and -out located at the train platform. The structure of the system means that the specific train used by passengers is unknown and needs to be estimated to obtain train load profiles [6]. To simplify the problem, we can use the subgroup of passengers called reference passengers [8]: a reference passenger is a passenger whose trip from an origin to a destination has a unique predominant path. By using reference passengers, the assignment problem is simplified since the possible vehicles used by the passengers can only exist in that particular direction of the route. Thereby ensuring that the tap-in and -out patterns can only belong to a vehicle on that route. We study the route from Aarhus H Station to Aalborg Station to ensure that the passengers traveling between these two points are reference passengers since there is only one railroad path connecting the two stations.

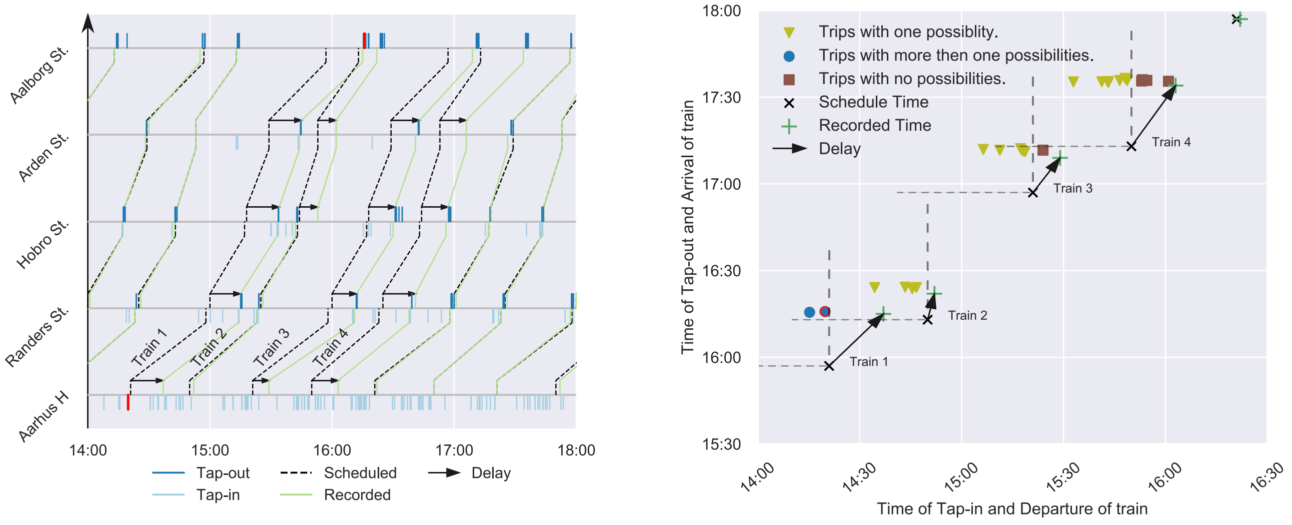
Using this route with reference passengers, we can investigate how the use of the scheduled timetable compared to the recorded timetable affects the trip to train assignment at different levels. The effects are investigated by applying the nearest neighbors algorithm with the assumptions that tap-in must happen before the departure and tap-out after the arrival of a train, giving three assignment approaches.

NN-TI: Assign trip-leg to the nearest departure of a train after tap-in occurred.
NN-TO: Assign trip-leg to the nearest arrival of a train before the tap-out occurred.
NN-TITO: Assign trip-leg to the nearest train, which departure occurred after the tap-in and which arrival occurred before tap-out.

With these approaches, we show how the tap-in and -out distributions are affected in Section II-B and how the assignments are affected in Section II-A. To have a ground truth for the assignment, we study the subset of ground truth trips by using the *NN-TITO* approach with the recorded time and keep the trips, where there is only one possible assignment. The assumptions are similar to more advanced Passenger-to-Train-Assignment methods [8], [19], [20], where the three approaches can be seen as how the weighting between tap-in and -out information affects the assignments.

A. ASSIGNING TRIPS TO TRAINS

To understand how the assignments of passengers are affected, we focus on a single passenger (Section II-A1)



(a) Trains from Aarhus to Aalborg on May 29, 2019 timespan 14.00–17.00. Red line shows the tap-in of a specific passenger at Aarhus station and the passengers tap-out at Aalborg Station.

(b) Entry-exit map of trips from Aarhus to Aalborg with time-span hour 14.00-18.00, on May 29, 2019. The trips are coloured by assignment possibilities using scheduled time, and the red passenger has a red circle around it.

FIGURE 1. **Subsection of trips tap-in and -out with trains arrival and departures on the May 29, 2019.**

initially, and then subsequently increase the number of passengers to display different ways the assignments are affected (Section II-A2) and lastly looking at the complete picture by studying the assignment of 49,458 trips to 15,136 trains spanning over three months (Section II-A3).

1) THE RED PASSENGER

Fig. 1(a) shows the tap-ins and -outs for departure and arrival of trains on the route from Aarhus to Aalborg Station in time-span hours 14.00 to 18.00 on May 29, 2019. An illustrative passenger traveling from Aarhus to Aalborg Station (tap-in at 14.20 and tap-out at 16.16) is highlighted in red. We see that trains 1 to 4 have notable delays as indicated by the black arrows, which show the change in time from the scheduled to the recorded time. The delay on the two first trains affects how the red passenger is assigned. If the *NN-TI* approach is used, then the red passenger will be assigned to train 1 regardless of using the scheduled or recorded timetable information. The red passenger using the *NN-TO* approach will be assigned to train 2 when using the scheduled timetable and to train 1 when using the recorded timetable. This is caused by the delay of the arrival of train 1 at Aalborg being so large that train 1 arrives after the scheduled arrival of train 2. For the *NN-TITO* approach, the red passenger using the recorded time is assigned to train 1, but when using the scheduled time, then train 1 and 2 are possible and will depend on the weighting of tap-in and tap-out to arrival and departure.

2) THE FOUR TRAINS

In total, there are 126 trips near the four trains in Fig. 1(a). Using the *NN-TI* approach, 87 of the trips will be assigned to the same train regardless of using recorded or scheduled departure time, and 39 trips will be assigned to different trains. Using the *NN-TO* approach, 16 of the trips will be assigned to a different train and 107 to the same train regardless of using recorded or scheduled arrival time. The majority of trips with different train assignment happens at Aalborg Station, where the delay of the train 1 is significant enough to surpass the scheduled arrival of train 2, which would assign all passengers to train 2. To see how the last approach affects assignments, we visualize the entry-exit map [13] of the four trains in Fig. 1(b). The figure shows the tap-ins and departure time of trains along the x-axis, and the tap-outs and arrival time of trains along the y-axis between Aarhus and Aalborg Station. The 31 trips in the figure can be divided into three groups using the scheduled departure and arrival time of the trains: 23 of the trips have only one possible assignment (yellow); 2 of the trips have more than one (blue) and 6 of the trips have no possibilities (red). Using the recorded time instead, will only give one trip with more than one assignment option, and the rest of the trips will only have one possible train assignment.

3) THE EFFECT OF SCHEDULED TIME ON TRAIN ASSIGNMENT

When comparing this ground truth with the scheduled time in combination with the three approaches, we see in Table 1 that the *NN-TI* and *NN-TO* approaches are both better than *NN-TITO*. The reason is that the *NN-TITO* approach inherits the errors from *NN-TI* and *NN-TO* in combination with their constraints making several trips impossible to assign to a single train. These constraints translate into a larger disagreement rate between the use of recorded and scheduled time. The large percentage point difference between the *NN-TI* and *NN-TO* is largely due to trains arriving early at a station, which create the peaks in the tap-out distribution of Fig. 2. If we take a subset of the ground truth and include

TABLE 1. Illustration of the difference between using scheduled timetable instead of the recorded for trip to train assignment for three different approaches. The assignment disagreement rate for an approach expresses the fraction of cases, where the assignment from scheduled and recorded time disagree. On the ground truth trips, all three approaches will give the correct assignment for recorded time.

The disagreement rate of using scheduled vs recorded time.

Approach	Ground truth trips[1)]	Ground truth trips with delay[2)]
NN-TI	3.86%	45.74%
NN-TO	22.19%	9.49%
NN-TITO	25.60%	48.97%

Fraction of ground truth trips affected by delay for each approach.

	Ground truth trips[1)]		Ground truth trips with delay[2)]	
Approach	Delay >0m	>10m	Delay >0m	>10m
NN-TI	40.80%	3.30%	100%	65.86%
NN-TO	55.17%	3.42%	100%	67.71%
NN-TITO	71.27%	4.30%	100%	80.48%

1) Contains 44548 trips with a single possibility using the third approach with recorded time. 2) Contains a subset of 2761 trips from the ground truth[1)], where there is a delay of 5 minutes on the arrival and departure time used.

the trains, which are a minimum of 10 minutes delayed, then the strength of the *NN-TO* emerges. The *NN-TI* approach is more sensitive to larger delays than the *NN-TO* since tap-ins are assigned to the next train when the tap-in is between the planned departure and recorded departure. For the *NN-TO* approach, the difference occurs when the arrival of the following train happens close to the previous train or trains arriving early. The difference between the different assignment rules depends on the patterns in delays and cancellation of trains.

B. EFFECT ON TAP-IN AND -OUT DISTRIBUTIONS

Fig. 2 shows the difference between using scheduled and recorded time when the train's arrival and departure time are not on time. Using the scheduled time at Aarhus Station for the tap-in, we see a thicker tail around 20-30 minutes before the departure and a sharp cut off at the departure time compared to recorded tap-in distribution. This happens when the passenger can catch the train due to a delay, but the tap-in is assigned to the next train since the scheduled time is used. The reason for the cut at 30 minutes before the departure is due to the headway of around 30 minutes at Aarhus Station. When using the recorded time for the tap-out at Aalborg Station, the tap-outs clustered together in one single peak by the first few minutes after the arrival of the train.

Compared to the tap-out distribution using the scheduled timetable, we see four peaks. The first peak is 30 seconds after the arrival, the seconds spread out around 4 minutes, and the next two are 15 and 45 minutes after the arrival. The second peak is due to delays, making it seem as if it takes a long time to tap-out. The three other peaks are due to the train arriving 1-2 minutes before the scheduled arrival time. When a train arrives before the scheduled time, the tap-out is assigned to the train before since the tap-out can only happen after the arrival of a train. In these cases, where the passengers are assigned to the previous train, the tap-out distribution will have peaks corresponding to the time-span between the arrival of trains. This is visible by the two last peaks around the headway of 15 and 45 minutes at Aalborg Station.

To mitigate the error of assigning passengers to the early arriving train, an *early slack* is usually used [6], which allows passengers to be assigned to trains a few minutes before the scheduled arrival. In this example, a 2 minutes *early slack* would be sensible and aligns with early arriving trains observed in Denmark [21].

III. MODEL

As discussed in the previous section, the arrival time of trains is robust for assigning passengers to trains using scheduled time with an *early slack*. At the same time, the departure can be inferred from the arrival time, since a train can only depart from a station after it has arrived [9]. The problem with using the scheduled arrival time with *early slack* is that it will give incorrectly tap-out distribution and will not tell which trains are on-time. To address these problems, we propose a hierarchical model, where the main observed variables are the tap-out time of reference passengers and the scheduled arrival times of the trains. In the model, we assume that the tap-outs for a given station on a given day are drawn from the same underlying distribution. Using the model, we can infer this tap-out distribution and infer the arrival times of trains. Importantly, it is possible to infer the arrival times, even where there are only a few passengers per train available. The following subsection is divided into the model steps, where the first subsection derives the model and defines its parts, the second step describes how to select the trains, which are informed by passengers, the third step describes how to infer the specific arrival time of the selected trains.

A. BAYESIAN MODEL

To infer the trains' arrival times, we propose a Bayesian model for the individual stations with $V \in \mathbb{N}$ vehicles, where each vehicle has a vehicle number $v \in [1, \dots, V]$. The observed variables are the tap-out times $\mathbf{T}^{\mathrm{O}} \in \mathbb{R}^N$ of $N \in \mathbb{N}$ passengers rides on a given station. The tap-outs are governed by a station specific walking behavior $\Omega \in \mathbb{R}^k$ of passengers aligning at the station. Given the tap-out times, we want to infer quantities representing the vehicle arrival times $\mathbf{A} \in \mathbb{R}^V$ measured in minutes since midnight, where the quantity $\mathbf{A}$ is a function of the scheduled arrival $\mathbf{A}^{\mathrm{sch}}$ and the delay $\boldsymbol{\delta} \in \mathbb{R}^V$, given by

$$\mathbf{A} = \mathbf{A}^{\mathrm{sch}} + \boldsymbol{\delta}. \tag{1}$$

Using Bayes' rule, we can write the posterior distribution of the arrival times and walking behavior given the tap-out

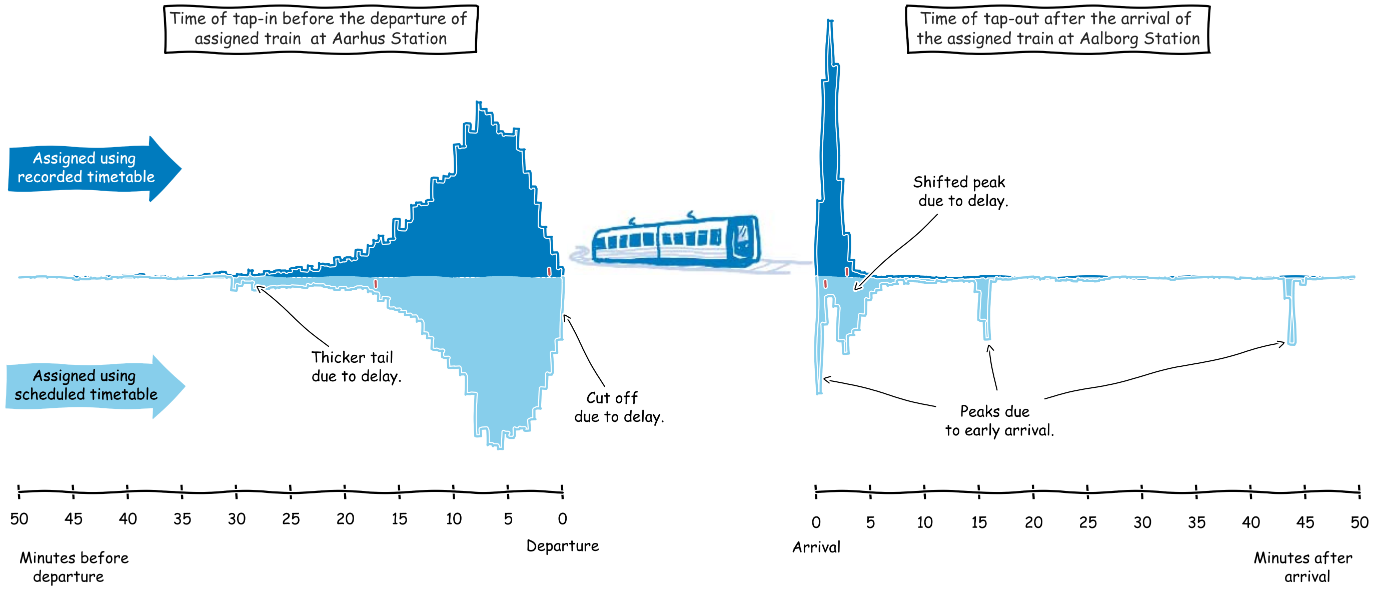


FIGURE 2. Illustration of the difference between using scheduled and recorded timetable on the ground truth trips tap-in and -out distribution (displayed as histograms) for trains not on-time from the complete data set. Assignment using recorded time is shown in dark blue, where scheduled time is shown in light blue. The axis indicates the minutes from the tap-in/-out to the time of the scheduled and recorded departure/arrival of the assigned train. As an example, a single passenger is highlighted in red to illustrate the effect.

times as

$$P\left(\mathbf{A},\Omega\middle|\mathbf{T}^{\mathrm{O}}\right)\propto P\left(\mathbf{T}^{\mathrm{O}}\middle|\mathbf{A},\Omega\right)P\left(\Omega\middle|\mathbf{A}\right)P(\mathbf{A}). \tag{2}$$

When deriving the model, we make the following assumptions regarding the vehicles arrivals and the passengers tap-outs times.

B. ASSUMPTIONS

Assumption (1) independence of stations: For simplicity and efficiency, we assume that the arrival time at one station is independent of the arrival time at the other stations. Clearly, this is a strong assumption, but it means that we will model and infer the arrival times for each station independently and in parallel using Eq. (2).

Assumption (2) independence of walking behavior and arrival time The walking behavior at a station is assumed to be independent of the arrival time such that

$$P\left(\Omega\middle|\mathbf{A}\right)=P(\Omega). \tag{A.3}$$

Assumption (3) independence of tap-outs: A given travelers tap-out time is assumed to be independent of other travelers tap-outs, when the arrival time of the trains are known, i.e.,

$$P\left(\mathbf{T}^{\mathrm{O}}\middle|\mathbf{A},\Omega\right)=\prod_{i}^{N}P\left(T_i^{\mathrm{O}}\middle|\mathbf{A},\Omega\right). \tag{A.4}$$

Assumption (4) conditional distribution of a tap-out: We assume that the conditional distribution of a tap-out $P(T_i^{\mathrm{O}}|A_{v_i},\Omega)$ only depends on the walking behavior and the arrival time of the ridden vehicle A_{v_i}, where v_i indicates the vehicle ridden by the i'th passenger. It is reasonable to assume that the conditional distribution of a tap-out is independent of the vehicle ridden v_i given the ridden vehicle's arrival time A_{v_i}, i.e., $P(T_i^{\mathrm{O}}|\mathbf{A},v_i,\Omega)=P(T_i^{\mathrm{O}}|A_{v_i},\Omega)$. Given this, we can write the joint conditional probability of tapping-out and the vehicle ridden by the passenger as

$$P\left(T_i^{\mathrm{O}},v_i\middle|\mathbf{A},\Omega\right)=P\left(T_i^{\mathrm{O}}\middle|A_{v_i},\Omega\right)P(v_i). \tag{A.5}$$

Assumption (5) sequence of ride events: When vehicle v arrives A_v, the passengers who rode the vehicle v_i can tap-out, i.e., a passenger can only tap-out T_i^{O} after the arrival of the vehicle ridden A_{v_i}, such that

$$A_{v_i}<T_i^{\mathrm{O}}. \tag{A.6}$$

Assumption (6) no overtaking: If the route has one track in a given direction and there are no overtake stop-points, then a given vehicle journey can not overtake the next vehicle journey unless the next vehicle journey is cancelled, which implies the following:

$$A_v<A_{v+1}. \tag{A.7}$$

C. THE FULL MODEL

Using assumption (A.3), the priors on walking behavior $P(\Omega)$ and arrival time are $P(\mathbf{A})$ independent. With the assumption (A.4) of conditional independence between tap-out times, the likelihood $P(\mathbf{T}^{\mathrm{O}}|\mathbf{A},\Omega)$ can be rewritten as the product of all tap-outs (8). The probability of each tap-out can be rewritten as the sum of all possible arrivals by summing over the different vehicles (9), then using assumption (A.5), the probability of tap-out i using vehicle v is independent of all other possible vehicles,

$$P\left(\mathbf{A},\Omega\middle|\mathbf{T}^{\mathrm{O}}\right)\propto\prod_{i=1}^{N}P\left(T_i^{\mathrm{O}}\middle|\mathbf{A},\Omega\right)P(\mathbf{A})P(\Omega) \tag{8}$$

$$=\prod_{i=1}^{N}\left[\sum_{v_i=1}^{V} P\left(T_i^{\mathrm{O}}, v_i \middle| \mathbf{A}, \Omega\right)\right] P(\mathbf{A})P(\Omega) \quad (9)$$

$$=\prod_{i=1}^{N}\left[\sum_{v_i=1}^{V} P\left(T_i^{\mathrm{O}} \middle| A_{v_i}, \Omega\right) P(v_i)\right] P(\mathbf{A})P(\Omega). \quad (10)$$

The final equation (10) has the form of a hierarchical mixture model with the likelihood $P(T_i^{\mathrm{O}}|A_{v_i}, \Omega)$ as mixture components, the mixture weights $P(v_i)$, and the priors $P(\mathbf{A})$ and $P(\Omega)$.

D. DEFINING THE MODEL PARTS

The mixture weights $P(v_i) = \theta_{v_i}$ is the prior probability of each vehicle and assumed to be Dirichlet distributed. The Dirichlet prior is a common prior used in mixture models [22], where we set the prior for all vehicles v to

$$\boldsymbol{\theta} \sim \text{Dirichlet}\left(\left[\frac{N}{V}\right]_{v=1}^{V}\right). \quad (11)$$

Setting the Dirichlet distribution with a hyperparameter equal to the ratio between the number of passengers and the number of trains correspond to a prior belief that more trains are utilized when there are more passengers.

The prior probabilities of the arrivals, $P(\mathbf{A})$, are specified through the relationship in Eq. (1) and a prior over the parameter δ_v that describes the delay of vehicle v from the scheduled arrival time A_v^{sch}. The distribution of delays can be described as having a high rate of trains with short delays, with a declining rate for trains with larger delay [21]. With this in mind, the delays are modeled using a truncated Student's-t distributed. The hyperparameters are set to give a high prior probability of small delays and a large spread. In addition, the delays are allowed to be negative of ℓ minutes capturing the *early slack*. The ℓ should be large enough to capture early arrivals but at the same time small enough not to capture the previous arrivals. i.e., we have

$$\delta_v \sim \text{Truncated Student-t}(\eta, \mu, \sigma_\delta, \ell) \quad (12)$$

where $\eta = 3$ is the degrees of freedom, $\mu = 0$ is the location parameter, $\sigma = 5$ is the scale parameter, and $\ell = -3$ based on the empirical observations in Section II-B.

The likelihood component $P(T_i^{\mathrm{O}}|A_{v_i}, \Omega)$ describes the tap-out distribution and depends on the walking time from the arrival of the train to card-reader on the given station. We assume that the tap-outs follow a Skew Generalized t-distribution (SGT) [23] with the behavior parameters $\Omega = (\sigma, \lambda, \mathrm{p}, \mathrm{q})$, such that

$$T_i^{\mathrm{O}} \sim \text{SGT}\left(A_v^{\text{sch}} + \delta_v, \sigma, \lambda, p, q\right). \quad (13)$$

All parameters are shared between all arrivals of vehicles except the parameters A_v^{sch} and δ_v controlling the location. By sharing the shape parameters σ, λ, p and q, the information from the tap-outs associated with the arrival of one train is used to inform the arrival of other trains.

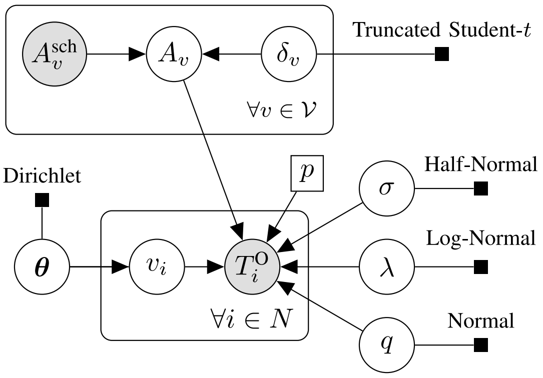


FIGURE 3. The model represented as a probabilistic graphical model.

The advantages of using the Skew Generalized t-distribution is that it can model a wide range of skewed distributions depending on its parameters, where p and q controls kurtosis, σ controls the scale, λ controls the skewness of the distribution. When we want to model the tap-out distribution, which can be described as a peaked skewed bell curve distribution with long tails [8] shown in Fig. 2, we need to constrain the parameters. Setting the constraint $0 < \lambda \leq 1$ for the skewness will make the distribution right-skewed, where fixing the $p = 2$ will ensure a bell curve distribution and $q > 0$ and $\sigma > 0$ ensures support for the distribution. The prior for q assumed to be a normal distribution with hyper-parameters puts weight on long tails for the tap-out distribution. The scale σ is assumed to be Truncated-normal with prior on a small variance for the tap-out since tap-outs tend to cluster together in a small interval [8], [9], [17]. Since the tap-outs are described as right-skewed distribution, we assume a normal distributing for the prior of the skewness λ with hyper-parameters imitated this. This means that the prior over the behavior parameters are

$$\sigma \sim \text{Truncated-Normal}(0.5, 0.25) \quad (14)$$

$$\lambda \sim \text{Truncated-Normal}(0.75, 0.5) \quad (15)$$

$$q \sim \text{Truncated-Normal}(2, 5) \quad (16)$$

The graphical structure of the model is shown in Fig. 3. The model structure can be seen as having two levels, where the first level is *vehicle specific level* with parameters θ and δ and the second level is the *behavioral level* with the parameters σ, λ, p and q of all vehicles.

E. IDENTIFYING TRAINS USED BY PASSENGERS

In some cases, the arriving trains will not have passengers tapping out with a smart card, making these inferred arrivals less informed. When an arrival has no passengers, the arrival time is inferred from the scheduled arrival time, encoded through the prior $P(\mathbf{A})$, the shared behavioral parameters and the assumption of no overtaking. To ensure that predictions are informed, we only make predictions for trains that are predicted to be used by passengers. The train used by the ith passenger is predicted as the component, i.e., the vehicle, with the highest marginal density for the passenger's tap-out

$$v_i^* = \left[\underset{v_i=1,\ldots,V}{\operatorname{argmax}} P\left(v_i \middle| T_i^{\mathrm{O}}\right) \right] \text{ for } i = 1, \ldots, N. \tag{17}$$

where the marginal distribution for the vehicle ridden v_i is

$$P\left(v_i|T_i^{\mathrm{O}}\right) = \iint P\left(T_i^{\mathrm{O}} \middle| A_{v_i}, \Omega\right) P(v_i)\, d\Omega\, dA_{v_i}. \tag{18}$$

The set of trains predicted to be used by passengers are then the *identified trains*

$$\mathcal{V}^* = \left\{ v_i^* \right\}_{i=1}^{N}. \tag{19}$$

F. DETERMINING THE ARRIVAL TIMES

When the subset of trains have been identified, the arrival time for each of the identified trains will be estimated using a three-step procedure. First, we predict the *approximately on-time* trains, which are defined as the set of trains, where the predicted delays are smaller than a threshold t. Second, if any tap-out predicted to use the *approximately on-time* train lies before the scheduled arrival time, it is reclassified to be an *early train*. In the third step, the *approximately on-time* trains are used to find the predicted percentile, which is the percentile of the posterior predictive distribution for the *approximately on-time* trains that minimizes the distance to the scheduled arrival time. To find the predicted *approximately on-time* trains, we calculate the difference between the samples from the posterior predictive of the identified trains and the scheduled arrival times A^{sch}. That is, we estimate the distribution of the differences between each combination of scheduled arrival times and the *identified trains* predictive posterior. These combinations can be used to calculate the probability d of the arrival time being in the range of t seconds of the scheduled arrival time, such that

$$\mathcal{V}^{\mathrm{sch}} = \left\{ \left(v^*, v\right) \in \mathcal{V}^* \times \mathcal{V} \mid P\left(|A_v^{\mathrm{sch}} - \widetilde{T}_{v^*}^{\mathrm{O}}| < t\right) > d \right\}, \tag{20}$$

where $\mathcal{V} = \{1, \ldots, V\}$ and $\widetilde{T}_{v^*}^{\mathrm{O}}$ follows the posterior predictive tap-out distribution for the v^*'th *identified train*

$$P\left(\widetilde{T}_{v^*}^{\mathrm{O}}|\mathbf{T}^{\mathrm{O}}\right) = \iint P\left(\widetilde{T}_{v^*}^{\mathrm{O}}|\mathbf{A}, \Omega\right) P\left(\mathbf{A}, \Omega, |\mathbf{T}^{\mathrm{O}}\right) d\Omega\, d\mathbf{A}. \tag{21}$$

There may be more than one of the *identified trains* or one of the scheduled arrival times that satisfy the condition in Eq. (20). For a non-unique match, we do not know if the train is on time or not. To mitigate this, we ensure unique matches between v^* and v by considering the set

$$\begin{aligned} \mathcal{V}^{\mathrm{sch}*} = \Big\{ \left(v^*, v\right) \in \mathcal{V}^{\mathrm{sch}} \mid \forall \left(\hat{v}^*, \hat{v}\right) \in \mathcal{V}^{\mathrm{sch}} \setminus \left\{ \left(v^*, v\right) \right\} : \\ \hat{v}^* \neq v^* \vee \hat{v} \neq v \Big\}. \end{aligned} \tag{22}$$

In addition, if any tap-out is predicted to have ridden the v'th vehicle lies before the associated scheduled arrival time, the arrival is classified to be an *early train*. This means that if we, for instance, use $d = 95\%$ and $t = 60$, the approximately on-time trains are defined as the subset of the *identified trains*, where at least 95% of the probability mass for the predictive tap-out distribution is within 1 minute of a unique scheduled arrival time, and all tap-out predicted to use the train lies after scheduled arrival time.

Therefore, in summary, the *approximately on-time* trains have two important characteristics. First, they are most likely used by at least one passenger because they are an *identified train*. Joining this with the knowledge that tap-outs cluster together right after the arrival of trains [9], we know that arrival time is likely near the *approximately on-time trains* predictive posterior. Secondly, they are predicted to be approximately on time since they are near a scheduled arrival time. In combination with the second characteristic, we can use the scheduled arrival time as a reference point to link distributions to the scheduled timetable and find which percentile in predictive posterior distribution, that is the best prediction for the arrival of the train. This percentile can then be used to identify the arrival time of the other trains. This is possible due to the shape of the predictive posterior coming from the behavioral level of the model, which is shared between all arrivals. Using *approximate on-time* trains, we can find the *predictive percentile p*, which minimizes the distance between the scheduled arrival time and predictive posterior of the *approximate on-time* train. However, if there is only one tap-out for a given *approximate on-time* train, the model will be very uncertain about the delay, and the percentile will not be representative of the arrival time. Therefore, we exclude trains with only one tap-out in the estimation of p^* (i.e., we remove them from $\mathcal{V}^{\mathrm{sch}*}$ before estimating p^*). Using Q as a percentile function, we have

$$p^* = \underset{(v^*, v) \in \mathcal{V}^{\mathrm{sch}*}}{\operatorname{Median}} \left(\underset{p}{\operatorname{argmin}} \left| A_v^{\mathrm{sch}} - Q\left(\widetilde{T}_{v^*}^{\mathrm{O}}, p\right) \right| \right). \tag{23}$$

After the predicted percentile p^* is found, the percentile is used to infer the arrival times of the full set of identified trains $\mathcal{V}^*$ predictive posteriors distribution $P(\widetilde{T}_{v^*}^{\mathrm{O}}|\mathbf{T}^{\mathrm{O}})$. If all trains are inferred to be delayed, then the day is classified as a *delayed day*. The predicted percentiles can be used from other inferred days of the same station to predict the *delay days*. In addition, if the goal is to recreate the complete recorded timetable, a train matching algorithm can be applied [9], [17], where the *approximately on-time trains* can be used as links to the scheduled timetable, thereby making it possible to assign the scheduled train IDs. The complete procedure for estimating the arrival time is described by pseudo-code in the Alg. 1.

IV. CASE STUDY

A. DATA SET

The study is conducted using data from AFC and AVL from the Danish company Rejsekort & Rejseplan A/S, which are the administrator of the national broad AFC system and the main travel planner in Denmark. The case study uses the subset of the route from Aarhus to Aalborg, where the trains stop at the following sequence of stations; Aarhus H, Hinnnerup,

Algorithm 1 Pseudo Code for Inferring the Arrival Time

```
Sample $\mathbf{A}, \Omega \sim P(\mathbf{A}, \Omega|\mathbf{T}^{\mathrm{O}})$ using HMC    ▷ Eq. 8–10
Determine the set of *identified trains* $\mathcal{V}^*$    ▷ Eq. 17–19
Match identified trains to time table    ▷ Eq. 20–22
**for** all matched trains **do**
    **if** there exists a tap-out before the scheduled time **then**
        Classify train as *early train*
    **else**
        Classify train as *approximately on-time*
Trains which are not matched are classified as *Not on-time*
**if** there are any *approximately on-time* trains **then**
    Estimate the percentile $p^*$    ▷ Eq. 23
**else**
    Classify the day as *delay day* and estimate the percentile $p^*$ from the previous days
**for** each *identified trains* $\mathcal{V}^*$ **do**
    Estimate arrival time using the $p^*$'th percentile of posterior predictive distribution $P(\widetilde{T}^{\mathrm{O}}_{v^*}|\mathbf{T}^{\mathrm{O}})$
```

TABLE 2. Origin destination matrix for reference passenger.

	Destination			
Origin	Randers	Hobro	Arden	Aalborg
Aarhus H	13991	4676	569	10139
Hinnerup St.	1	3	-	6
Hadsten St.	712	324	9	523
Langå St.	1462	238	48	774
Randers St.	-	1999	410	3951
Hobro St.	-	-	1156	8208
Arden St.	-	-	-	2734
Total	16166	7240	2192	26335

Hadsten, Langå, Randers, Hobro, Arden and Aalborg. The stations Randers, Hobro, Arden and Aalborg vary in size and activity. Therefore, these are chosen to investigate the behavior of the model for different sizes of the stations. The number of trip-legs traveling to and from these stations is shown in Table 2. The data are from the period February 1st to May 31st 2019, excluding days where the recorded timetable information was not available. The complete data set contains the last trip leg of 51,933 reference passengers, and 15,136 intercity trains arrivals for the four stations.

The difference between the scheduled and the recorded arrival time can be zero, positive or negative [21] creating a timetable error (*TT error*). There are 56% of the arriving trains, which have a non-zero TT error. When there is a difference, the majority of the errors is below 2 minutes, amounting to 25% of the trains. Table 3 shows summary statistics for the *TT error* for each station. Since a large portion of the trains has no or a small *TT error*, the mean *TT error* of the four stations is around 1 minute for each station shown in the table. The larger stations Hobro and Aalborg are the stations with the largest share of small delays, where the greater delays are more evenly spread between the stations. The reason behind this is that minor *TT errors* can easily be minimized between stations, where greater *TT errors* will persist in the subsequent stations even if some time is gained.

TABLE 3. The share of trains having a difference between recorded and scheduled arrival and descriptive statistics of the difference.

	Randers St.	Hobro St.	Arden St.	Aalborg St.
On-time	39.88%	68.37%	67.59%	49.61%
$0 <$ TT error < 2	32.18%	18.33%	19.18%	30.27%
$2 \leq$ TT error < 10	23.24%	9.17%	8.43%	15.88%
$10 \leq$ TT error	4.70%	4.13%	4.80%	4.24%
Descriptive statistics in minutes				
Mean	1.86	1.18	1.19	1.55
Std	4.14	4.22	4.05	4.29
Max	52.00	68.00	62.00	62.00

Overview of the implementation and section: The model is implemented in the probabilistic programming language STAN [24] using the Hamiltonian Monte Carlo with No U-turn sampler [25] with four chains of minimum 6000 iterations and warm-up period of 4000. Mixture models are known for label switching [26] and relabeling is done within and between chains. For each station, the arrival times are inferred by first selecting the identified trains for arrival, which is discussed in Section IV-B. The learning and the uncertainty of the *identified trains* arrival time are discussed in Section IV-C, while the results are presented and evaluated in the last Section IV-E. As described in assumption 1, the model is fitted separately to data for each day and each station.

B. SELECTION OF THE IDENTIFIED TRAIN

The travel activity varies on the different stations, and not all arriving trains will have passengers tapping out with a smart card. We can determine the *identified trains*, which are most likely used by the passengers, using the selection procedure described in Section III-E. In Table 4 we see that the number of *identified trains* arrivals is positively correlated with the number of passengers going to the given station. At Aalborg station, 93% of train arrivals are identified with an average activity of 201 trips per day. In contrast, at Arden station, 29% of train arrivals are identified with an average of 16 trips per day.

The difference in the share of *identified train* arrivals between stations originates from how the passengers are distributed among the trains during the day. Fig. 5 shows the tap-outs during the day, with the recorded train arrivals and predictive posterior of equation (21). In Fig. 5(b) for Aalborg station all recorded arrivals have a tap-out nearby except for the morning train 1, which translates into a lower density peak for the posterior than the rest of the day. For the case of Arden station in Fig. 5(a) the pattern is more apparent, where only 11 out of the 37 trains have a tap-out nearby. The train arrivals with tap-outs nearby are transformed into

TABLE 4. **Identified train arrivals, average trips and trains.**

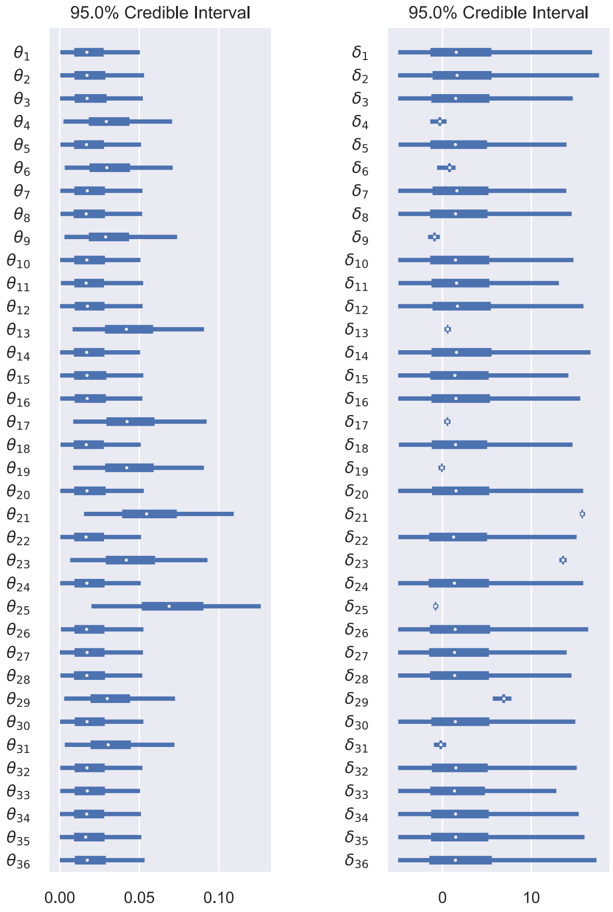
		Randers	Hobro.	Arden	Aalborg
Trains	Identified	3270	2626	1083	3497
	% of total	86.90%	69.78%	28.78%	92.93%
Average pr. day	Trips	126	56	16	201
	Trains	35	35	35	35
Trips per train & day	Min	1.941	0.972	0.118	3.824
	Mean	3.619	1.612	0.457	5.734
	Max	6.207	3.528	0.833	10.750
Trips assign per Identified train	Min	1.000	1.000	1.000	1.000
	Mean	4.079	2.295	1.508	5.992
	Max	20.000	22.000	6.000	62.000

well-defined peaks of the predictive posterior with low variance and higher density. The density and variance reflect the uncertainty of the inferred train arrivals. Since the only vehicle-specific parameters are the delay δ_v, and the probability of the vehicle θ, the difference in variance and density of the predictive posterior peaks stems from these parameters. This relation is clearer when comparing the position of the peaks in Fig. 5(a) with the corresponding index of the posterior distribution of θ and δ in Fig. 4. The peaks with nearby tap-outs correspond to the θ's with higher means and the δ's with small variance compared with those without tap-outs. The two θ's with the highest mean values are the 21st and 25th, which are the two train arrivals with most passengers on that given day.

C. UNCERTAINTY OF THE IDENTIFIED TRAIN ARRIVAL TIME

After the *identified trains* are found, the predictive posterior of *identified trains* $P(\widetilde{A}_{v*}|\mathbf{T}^{\mathrm{O}})$ can be used to predict the train arrival times at the station. As described in Section III-F the *approximate on-time trains* are used to determine the *predicted percentile*. The *predicted percentile* will vary from day to day and station to station. This difference originates from different behavior at the stations and the different levels of available information at the stations, given different degrees of uncertainty. Informally, the uncertainty of the *identified trains* predictive posterior is due to the small number of passengers at the given station during the day and the diffuse domain knowledge. In the Bayesian framework, the prior distribution captures our domain knowledge of the parameters before we observed the data. The posterior distribution summarizes our knowledge after observing the data and can be seen as a compromise between the prior distribution and the likelihood.

In order to understand how this affects the model, it is important to understand how the model learns its parameters. As mentioned in Section III-D, the model has two levels: *the vehicle specific level* and *the behavioral level*. Prior assumptions about each level are encoded by the respective priors. The likelihood is informed by each tap-out introduced to the model, and the product of the prior and likelihood creates the posterior distribution. However, the degree of information obtained by the likelihood will vary differently at the *the vehicle specific level* (Δ) and *the behavioral level* (Ω) of the model.

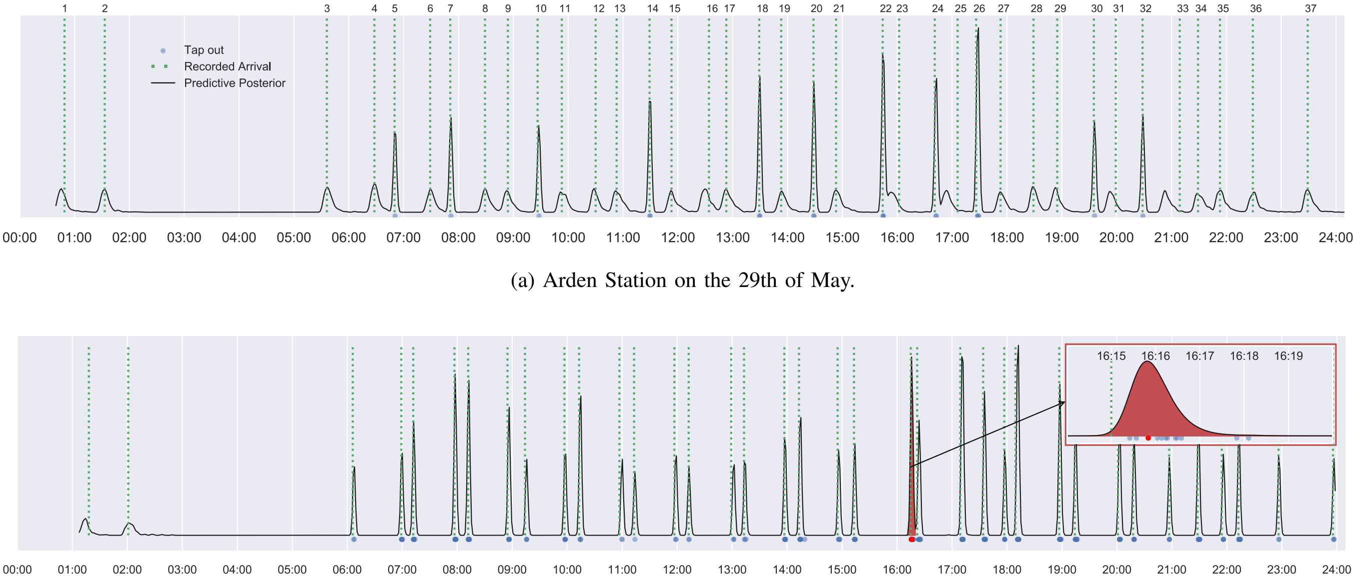


(a) Posterior of θ's

(b) Posterior of δ's

FIGURE 4. **The posterior distribution of the parameters δ and θ at Arden Station on the 29th of May 2019.**

To understand this, we can see the priors as regularization that control how likely some values are relative to others and the initial uncertainty of the parameter. Weaker priors will make a broader range of values more likely, thereby being more uncertain about the true value vice versa. In the case of the prior on the vehicle arrivals time $\mathbf{A}$, the prior is weakly informative because of the prior on the delay parameter δ, where the probability density covers a large area from small to large values. At the same time, we know that the train tap-outs tend to cluster together [9], meaning that the likelihood will put a larger weight on a small area of possible δ-values. Combining the likelihood with the prior, the posterior will put a large amount of its density at the same area as the likelihood since the weighting is so strong compared to the prior belief, even when there are only a few tap-outs. The priors of the walking behavior Ω are more informed, which means that the model needs more evidence for the posterior distribution to change from the prior belief through the likelihood.



(a) Arden Station on the 29th of May.

(b) Aalborg Station on the 29th of May.

FIGURE 5. **Tap-outs during the day with the actual train arrivals and the predictive posterior for Aalborg and Arden stations. The predictive posterior distributions from equation (21) as a function of time is shown as a black line. The actual arrival times are indicated as a green dotted line and with tap-outs as shaded blue. A part of the predictive posterior distribution is enlarged (shaded red), where it is possible to see the tap-out is more spread out. The red dot is the red passenger tapping-out at 16.16 in hours, who is predicted to belong to this component.**

Suppose we simplify the model to two components, where there are two arrivals, each with two associated tap-outs. The first arrival is a few minutes delayed, and the second has a considerable delay. Then the model has three possible ways of getting the density to concentrate near the actual train arrival times and tap-outs, i.e., increasing the likelihood of the data: either changing the shape through Ω, changing the position through $\boldsymbol{\delta}$ or both. Since the prior on $\boldsymbol{\delta}$ is weaker than the prior on Ω, the model is more likely to change to position than the shape of the posterior distribution to increase the likelihood.

The mean of each component will stabilize around the tap-outs, and the tap-outs near a component will be informative for the associated δ. The number of passengers tapping out near the arrival of a train gives more information on the degree of delay for the specific train, thereby a lower uncertainty for the delay δ. Thereby the uncertainty of $\boldsymbol{\delta}$ is depending on the number of passengers tapping out near the arrival of a train. In the cases from Fig. 5, Aalborg station has an average of 6.8 per passengers per *identified train* resulting in an average standard deviation of 0.077 for the associated delays, where Arden station has 2.0 passengers per *identified train* with an average standard deviation of 0.219 for the associated delays.

More information is required to decrease the posterior uncertainty of the behavioral parameters $\Omega = (\sigma, \lambda, q)$ than for the delay $\boldsymbol{\delta}$. However, the shared nature of *the behavioral level* means that the tap-outs from all components are used to infer the behavioral parameters. In our simple example from above, each delay δ will mainly be inferred by the two tap-outs near the component, where all four tap-outs will inform the behavioral parameters. If we add a component with five tap-outs, then the uncertainty of Ω will decrease even further by being informed by nine tap-outs. This means that the uncertainty of Ω depends on the total number of tap-outs there are during the day.

TABLE 5. **The test set RMSE for the Bayesian model, S-Epoch and Denclue method**

		Randers St.	Hobro St.	Arden St.	Aalborg St.
Overall	Bayes	0.74	0.69	0.59	0.90
	Bayes++	0.67	0.64	0.58	0.75
	Denclue	1.10	1.00	0.96	1.25
	Epoch	0.97	0.84	0.81	1.24
Early	Bayes	0.84	0.59	0.52	0.61
	Bayes++	0.80	0.59	0.50	0.53
	Denclue	1.09	0.66	0.71	1.09
	Epoch	0.93	0.80	0.63	1.09
On time	Bayes	0.64	0.62	0.66	0.91
	Bayes++	0.57	0.58	0.69	0.78
	Denclue	1.21	1.02	1.34	1.26
	Epoch	0.98	0.84	1.08	1.36
Late	Bayes	0.74	1.02	0.71	1.11
	Bayes++	0.63	0.91	0.63	0.89
	Denclue	0.97	1.35	0.73	1.39
	Epoch	1.02	0.89	0.79	1.19

The table is divided into an overall result and the results depending on the observed train arrivals with the three classifications "Early", "On time" and "Late". The "Bayes++" uses a shift instead of the percentile (section III-F) for estimating the trains arrival with only one tap-out.

D. CROSS VALIDATION AND EVALUATION METRICS

The predicted arrival time will vary in performance from day to day and station to station depending on the degree of delay and the choice of the probability d of the arrival

TABLE 6. Classification of the arrival for the Bayesian models test set.

	Prediction		Randers St.	Hobro St.	Arden St.	Aalborg St.
Early (34.49%)	Early train	Model error	0.78	0.57	0.52	0.61
		TT error	2.37	1.63	1.52	1.76
		Fraction	89.88%	96.81%	99.12%	99.4%
	Approx On Time	Model error	1.10	0.96	0.69	0.89
		TT error	1.39	1.35	1.09	1.11
		Fraction	9.37%	2.88%	0.88%	0.5%
	Not On Time	Model error	2.10	1.07	- (-)	1.30
		TT error	1.35	1.80	- (-)	2.08
		Fraction	0.75%	0.32%	0.00%	0.10%
On Time (44.68%)	Early train	Model error	3.07	2.21	0.54	2.06
		TT error	0.00	0.00	0.00	0.00
		Fraction	0.27%	0.28%	1.72%	0.46%
	Approx On Time	Model error	0.46	0.46	0.35	0.59
		TT error	0.00	0.00	0.00	0.00
		Fraction	81.84%	90.78%	93.47%	68.84%
	Not On Time	Model error	1.09	1.42	2.59	1.37
		TT error	0.00	0.00	0.00	0.00
		Fraction	17.89%	8.94%	4.47%	30.7%
	Late Day	Model error	- (-)	- (-)	2.26	- (-)
		TT error	- (-)	- (-)	0.00	- (-)
		Fraction	0.00%	0.00%	0.34%	0.00%
Late (20.82%)	Early train	Model error	0.11	0.56	- (-)	3.31
		TT error	2.00	6.00	- (-)	6.78
		Fraction	0.13%	0.28%	0.00%	0.63%
	Approx On Time	Model error	0.75	0.48	0.47	1.18
		TT error	1.00	1.01	1.00	1.00
		Fraction	10.4%	10.26%	21.5%	1.88%
	Not On Time	Model error	0.74	1.06	0.75	1.08
		TT error	7.87	10.33	9.46	7.54
		Fraction	89.47%	89.46%	71.96%	97.49%
	Late Day	Model error	- (-)	- (-)	0.82	- (-)
		TT error	- (-)	- (-)	11.59	- (-)
		Fraction	0.00%	0.00%	6.54%	0.00%

The first column divided the observed train arrivals into three classifications "Early", "On time" and "Late" with the following parenthesis showing the fractions of trains in each class. Each of the three classes is subdivided into four predictions ("Early Train", "Approx on time", "Not on time" and "Late day"). For each prediction and station, the model error and time table error ("TT error") are shown as RMSE compared to the recorded time and measured in minutes. The fraction of trains ("Fraction") within each of the four predictions are also shown for each class and station.

TABLE 7. Chosen probability *d* of the arrival time being in the range of *t* seconds of the scheduled arrival time.

		Randers St.	Hobro St.	Arden St.	Aalborg St.
Most frequent	$(t, d)^{1)}$	(60,10)	(80,10)	(160,95)	(60,5)
	Rate	100.0%	77.07%	95.77%	100.0%
Average	t	60.00 (0.00)	75.52 (8.76)	158.77 (6.03)	60.00 (0.00)
	d	10.00 (0.00)	11.25 (3.14)	94.20 (4.47)	5.00 (0.00)
	p^*	1.96% (2.57)	5.11% (5.55)	15.33% (10.66)	1.64% (1.88)

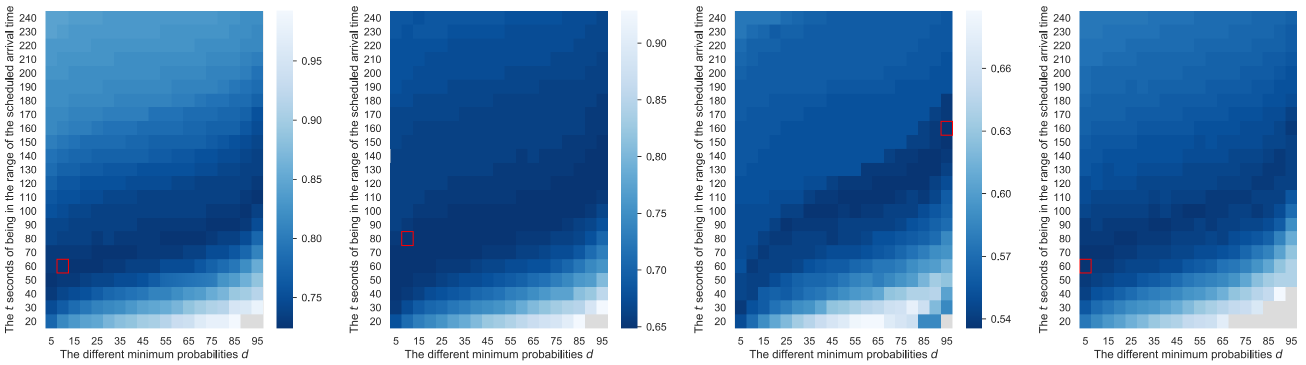
Round parenthesis indicates standard deviation. 1) The combinations of (t, d) are chosen using grid search on training set with values $t \in [20, \ldots, 240]$ and $d \in [5\%, \ldots, 95\%]$.

time being in the range of t seconds of the scheduled arrival time. Leave-one-out cross-validation is performed to assess how well the model generalizes, where each fold of the test set is equal to a day. On the training set, the Bayesian model first finds the *identified trains* for each day and station, then selects the combination of (t, d), which minimizes the Root Mean Square Error (RMSE) in the training set. The chosen combination of (t, d) from the training set is then used on the test-set to give the test-set RMSE. The RMSE is taken between the predicted arrival time and the actual arrival time, where the combination of (t, d) is found using grid-search with the values $t \in [20, \ldots, 240]$ and $d \in [5\%, \ldots, 95\%]$. For each combination of (t, d), the model will classify the estimated trains arrivals into *early*, *approximate on time*, *not on time* and *late days*. These classifications can then be compared to actual arrivals by dividing them into early, on-time or late trains, depending on the actual arrival time being before, the same or after the scheduled arrival time and evaluated by the fraction of correct and incorrect classifications.

E. THE PREDICTED PERCENTILE AND THE RESULTING ARRIVAL TIME

In Table 5, the Bayesian model is compared to the S-Epoch method [9] and the *kernel density estimation* based method Denclue [18] using cross-validation. The parameters for both methods are optimized to minimize the RMSE of the training set and used on the test set. Overall the Bayesian model outperforms the two other models with a lower RMSE. If we consider the division of the results into the classifications of the observed delays as "Early," "On time" and "Late," the pattern persists, except for the late trains at Hobro St. The Bayesian model can be further improved by using a shift (similar to S-Epoch and Denclue) instead of the percentile (Section III-F) for estimating train arrivals with only one tap-out. The improvement is denoted by "Bayes++" in Table 5.

In Table 6, the model's classification for the test set is compared with observed classifications. We see that the model is efficient in classifying the *early trains* as *early* with the lowest fraction of correct classification of 89% for Randers Station. Notable the model errors are around half of the TT errors for this classification. The observed on-time trains is a more mixed picture, where the Arden Station have the highest correct classified with *Approximate on-time* having a fraction of 93%, where the Aalborg Station has the lowest with a fraction of 68%. However, Aalborg station has the highest correctly classified fraction of the late trains as *Not On Time* with 97%, where Arden Station have the lowest correctly classified fraction of 71%. This illustrates the challenge the model can have with minor delays, where it has difficulty separating the trains with delay and without. Despite this, the trains with larger delays are classified as *not on time*. The model can also make passenger to train assignments since the identified trains are the trains predicted to be used by passengers. Applying equation (17) to predicting



(a) Randers Station. (b) Hobro Station. (c) Arden Station. (d) Aalborg Station.

FIGURE 6. The average RMSE for the test set for all (t, d) combinations. The (t, d) combination that is selected as optimal on the training set is indicated as a red square. Grey areas are combinations of (t, d) that have non approximated on-time trains.

the train used by each tap-out gives a hit rate of 99.54% with the ground truth trips. In general, when the model has a higher fraction of incorrect classifications, the model error is around 1-3, where the correct classification of the model error is around 0.5 to 1.0 RMSE.

In Table 7 the chosen combination of (t, d) is stable for Randers and Aalborg Station, having only one combination chosen for each station, where Hobro and Arden Station have more variation with the most frequent combination chosen 77% and 95% of the time, respectively. This is more clear when looking at Fig. 6, where Randers and Aalborg Station have a tighter optimal than the stations Arden and Hobro. All the station see clear valleys of combination, where the combination is optimal, indicated by dark blue area.

V. CONCLUSION

In this paper, we have shown how erroneous timetable information can affect the inference of tap-in and -out distribution and how the assignment of passengers to trains is affected. To diminish the error, the paper proposes a hierarchical Bayesian model to infer the arrival time of trains, where the only input is the scheduled arrival time and tap-outs from a single station. The results show that the model can infer 70% of arrival times with an average error of 28 to 32 seconds. Since the model assumes no-overtaking, the predicted arrival times of the model can easily be matched with the ID of the correct trains. In cases where this assumption does not hold, it would be an interesting research problem to extend the model to handle overtaking. In addition to this, further directions of research would be to challenge the model's main limitations regarding the assumption of independence of stations, how to gain more information from inferred components with a single tap-out and how better to distinguish trains with minor delays from trains without delay.

ACKNOWLEDGMENT

The authors would like to thank Rejsekort & Rejseplan A/S for providing the data for the project.

REFERENCES

[1] C. Iliopoulou and K. Kepaptsoglou, “Combining ITS and optimization in public transportation planning: State of the art and future research paths,” *Eur. Transp. Res. Rev.*, vol. 11, no. 1, pp. 1–27, 2019.

[2] K. E. Zannat and C. F. Choudhury, “Emerging big data sources for public transport planning: A systematic review on current state of art and future research directions,” *J. Indian Inst. Sci.*, vol. 99, pp. 601–619, Oct. 2019.

[3] M. P. Pelletier, M. Trépanier, and C. Morency, “Smart card data use in public transit: A literature review,” *Transp. Res. C Emerg. Technol.*, vol. 19, no. 4, pp. 557–568, 2011, doi: 10.1016/j.trc.2010.12.003.

[4] S. Robinson, B. Narayanan, N. Toh, and F. Pereira, “Methods for pre-processing smartcard data to improve data quality,” *Transp. Res. C Emerg. Technol.*, vol. 49, pp. 43–58, Dec. 2014, doi: 10.1016/j.trc.2014.10.006.

[5] S. Robinson and M. Manela, “Automatic identification of vehicles with faulty automatic vehicle location and control units in London buses' iBus system,” *Transp. Res. Rec.*, vol. 2277, no. 1, no. 2277, pp. 21–28, 2012. [Online]. Available: https://doi.org/10.3141/2277-03

[6] D. Luo, L. Bonnetain, O. Cats, and H. van Lint, “Constructing spatiotemporal load profiles of transit vehicles with multiple data sources,” *Transp. Res. Rec. J. Transp. Res. Board*, vol. 2672, no. 8, pp. 175–186, 2018.

[7] A. Tavassoli, A. Alsger, M. Hickman, and M. Mesbah, “How close the models are to the reality? Comparison of transit origin-destination estimates with automatic fare collection data,” in *Proc. Aust. Transp. Res. Forum (ATRF)*, 2016, pp. 1–6.

[8] S. P. Hong, Y. H. Min, M. J. Park, K. M. Kim, and S. M. Oh, “Precise estimation of connections of metro passengers from smart card data,” *Transportation*, vol. 43, no. 5, pp. 749–769, 2016, doi: 10.1007/s11116-015-9617-y.

[9] Y. H. Min, S. J. Ko, K. M. Kim, and S. P. Hong, “Mining missing train logs from smart card data,” *Transp. Res. C Emerg. Technol.*, vol. 63, pp. 170–181, Feb. 2016, doi: 10.1016/j.trc.2015.11.015.

[10] A. Alsger, B. Assemi, M. Mesbah, and L. Ferreira, “Validating and improving public transport origin-destination estimation algorithm using smart card fare data,” *Transp. Res. C Emerg. Technol.*, vol. 68, pp. 490–506, Jul. 2016, doi: 10.1016/j.trc.2016.05.004.

[11] N. Nassir, M. Hickman, and Z. L. Ma, “Activity detection and transfer identification for public transit fare card data,” *Transportation*, vol. 42, pp. 683–705, Apr. 2015.

[12] F. Zhou and R. H. Xu, “Model of passenger flow assignmentfor Urban rail transit based on entryand exit time constraints,” *Transp. Res. Rec.*, vol. 2284, no. 1, pp. 57–61, 2012.

[13] T. Kusakabe, T. Iryo, and Y. Asakura, “Estimation method for railway passengers' train choice behavior with smart card transaction data,” *Transportation*, vol. 37, no. 5, pp. 731–749, 2010.

[14] A. M. Wahaballa, F. Kurauchi, T. Yamamoto, and J. D. Schmöcker, “Estimation of platform waiting time distribution considering service reliability based on smart card data and performance reports,” *Transp. Res. Rec.*, vol. 2652, no. 1, pp. 30–38, 2017.

[15] A. Tavassoli, M. Mesbah, and A. Shobeirinejad, “Modelling passenger waiting time using large-scale automatic fare collection data: An Australian case study,” *Transp. Res. F Traffic Psychol. Behav.*, vol. 58, pp. 500–510, Oct. 2018.

[16] P. Kumar, A. Khani, and Q. He, “A robust method for estimating transit passenger trajectories using automated data,” *Transp. Res. C Emerg. Technol.*, vol. 95, pp. 731–747, Aug. 2018. [Online]. Available: https://doi.org/10.1016/j.trc.2018.08.006

[17] H. E. Tan, D. W. Soh, Y. S. Soh, and M. A. Ramli, “Derivation of train arrival timings through correlations from individual passenger farecard data,” *Transportation*, to be published.

[18] A. Hinneburg and H.-H. Gabriel, “DENCLUE 2.0: Fast clustering based on kernel density estimation,” in *Advances in Intelligent Data Analysis VII*, M. R. Berthold, J. Shawe-Taylor, and N. Lavrač, Eds. Berlin, Germany: Springer, 2007, pp. 70–80.

[19] Y. Zhu, H. N. Koutsopoulos, and N. H. Wilson, “A probabilistic passenger-to-train assignment model based on automated data,” *Transp. Res. B Methodol.*, vol. 104, pp. 522–542, Oct. 2017, doi: 10.1016/j.trb.2017.04.012.

[20] W. Zhu, W. Wang, and Z. Huang, “Estimating train choices of rail transit passengers with real timetable and automatic fare collection data,” *J. Adv. Transp.*, vol. 2017, Aug. 2017, Art. no. 5824051.

[21] F. Cerreto, B. F. Nielsen, O. A. Nielsen, and S. S. Harrod, “Application of data clustering to railway delay pattern recognition,” *J. Adv. Transp.*, vol. 2018, Apr. 2018, Art. no. 6164534.

[22] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis*, 3rd ed. London, U.K.: Chapman & Hall, 2014, ch. 22.4, p. 675. [Online]. Available: http://www.stat.columbia.edu/ gelman/book/

[23] P. Theodossiou, “Financial data and the skewed generalized t distribution,” *Manag. Sci.*, vol. 44, no. 12, pp. 1650–1661, 1998.

[24] B. Carpenter *et al.*, “STAN: A probabilistic programming language,” *J. Stat. Softw.*, vol. 76, no. 1, pp. 1–32, 2017.

[25] M. D. Hoffman and A. Gelman, “The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo,” *J. Mach. Learn. Res.*, vol. 15, pp. 1593–1623, Nov. 2014. [Online]. Available: http://jmlr.org/papers/v15/hoffman14a.html

[26] P. Papastamoulis, “Label.Switching: An R package for dealing with the label switching problem in MCMC outputs,” *J. Stat. Softw.*, vol. 69, no. 1, pp. 1–33, 2016.

Chapter 3

Paper B - Estimation of transfer walking time distribution in multimodal public transport systems based on smart card data*

The following pages contain the article which has been published in Transportation Research Part C: Emerging Technologies, vol. 132, 2021, doi: 10.1016/j.trc.2021.103332.*

*Equal contribution by authors, see author statements.

Estimation of transfer walking time distribution in multimodal public transport systems based on smart card data

Morten Eltved[a,1,*], Philip Lemaitre[b,1], Niklas Christoffer Petersen[a,1]

[a]Department of Technology, Management and Economics
Technical University of Denmark
Bygningstorvet 116B
Kgs. Lyngby, Denmark
[b]Department of Computer Science
IT University of Copenhagen
Rued Langgaards Vej 7,
Copenhagen, Denmark

Abstract

Transfers are a major contributor to travel time unreliability for journeys in public transport. Thus, connections between services in the public transport network must be reliable. To plan such reliable transfers from e.g. busses to trains, it is crucial to know the necessary walking times from stops to platforms. This paper presents an innovative approach for estimation of walking time distributions from bus stops to train platforms based on a matching of smart card data and automatic vehicle location data. The observed times from bus stop to rail platform turns out to have a large variance, due to two reasons: differences in passenger walking speeds, and passengers who are doing activities during the transfer. To account for these variations a *hierarchical Bayesian mixture model* is applied, where the time for passengers walking directly and passengers doing activities during the transfer follows separate distributions. The proposed methodology is applied to 129 stations in the Eastern part of Denmark, where the tap-in devices are located at the train platform. Results from two stations with different characteristics are presented in details along with justifications and analyses of model accuracy. The outcome of the model with distributions of the necessary walking times from bus stops to train platforms is important input for timetabling connections, and the data-driven methodology can easily be applied at scale.

Keywords: Public transport, Transfers, Walking time, Smart Card, Automatic Vehicle Location

[*] Morten Eltved: *morel@dtu.dk*

[1] *Authors are in alphabetical order.*

1. Introduction

The attractiveness of public transport is defined by many parameters, but transfers between services are consistently viewed as inconvenient (Iseki and Taylor, 2009; Raveau et al., 2014; Schakenbos et al., 2016). Transfers require the passenger to alight a service, and in most cases walk to another stop to board the connecting service. When transferring between services there is a risk of a large increase in the journey time of the whole trip if a connecting service is missed (Dixit et al., 2019), and thereby decreasing the reliability of the trip, which is known to be of large nuisance to passengers (Kouwenhoven et al., 2014).

Creating good connections between services require knowledge on the time needed for passengers to walk from one stop to another (Parbo et al., 2014). This knowledge is usually determined by identifying the walkways between stops and assuming a walking speed for the passengers, or by manual surveys where passengers are followed through the station (Daamen et al., 2006). If walking times are overestimated, this would lead to high waiting times at coordinated transfers, while underestimated walking times would lead to passengers missing planned connections and thus impose a large increase in the total travel time of the passengers. Planning connections between services thus rely on accurate measures of the needed walking times at transfers. Recent developments within timetable planning are able to incorporate the uncertainties of walking times and vehicle travel times, making it an important task to provide estimations on the necessary walking times at transfers (Xiao et al., 2016).

This paper presents a novel methodology for estimating the walking time distribution for transferring passengers from busses to train stations. The study utilises the vast available amount of automatic fare collection (AFC) data from smart cards and combines this with automatic vehicle location (AVL) data from busses. In this way it is possible to calculate the walking time for passengers from alighting at the bus stop until the passenger taps in at a validator device on the platform. However, the raw data can not be used directly for estimation of the required walking time, since passengers may be doing activities during their transfers (Wahaballa et al., 2018). A *hierarchical Bayesian mixture model* with one distribution for passengers walking directly and another distribution for passengers having an activity during the transfer is estimated, to obtain accurate estimates of the walking time distribution for directly walking passengers. The method is applied to a large scale case study and results are studied in detail for two stations with different characteristics.

The novel methodology adds to existing knowledge of transferring passengers by separating passengers walking directly and passengers doing activities during the transfer, and does this using an unsupervised method. The approach is able to handle different types of transfers, where either the synchronisation of busses and trains or the number of shops near the station increases the amount of activities undertaken by passengers during the transfer. The methodology can be easily applied at scale, and thus overcomes the scalability issues of time consuming manual surveys where passengers are followed through the station.

The paper is organised in the following way; Section 2 reviews the existing studies on estimation of walking times at transfers, Section 3 outlines the methodology for estimation of walking times based on smart card data, Section 4 presents the case study used for testing the methodology and analyses of the results, Section 5 discusses the model accuracy with possible verification techniques that can be applied at scale. Finally, Section 6 concludes on the findings in the paper.

2. Literature review

Walking is a central part of using public transport, and in many cases the passenger also needs to walk due to a transfer between services. The number of trips in metropolitan areas requiring a transfer can range between anywhere from 30 % to 80 % depending on the network layout and which modes of public transport the passengers use (Guo and Wilson, 2011). For the Greater Copenhagen area, which is part of the area used for the case study presented in Section 4, the number of trips requiring at least one transfer is approximately 40 % (Anderson, 2013). Given the large number of transfers in the network, it is important to estimate the necessary walking times for these transfers to achieve good coordination between buses and trains.

Walking speeds are known to be heterogeneous (Fruin, 1971), even when there is nothing that constrains the walkways (Daamen and Hoogendoorn, 2006). A number of studies have spent significant efforts for obtaining walking times at different transport facilities. Young (1999) for example studies the walking speeds in airport terminals and find that moving walkways and passing obstructions in a corridor significantly impact the walking speed. For public transport stations, Chen et al. (2016) studies the walking speeds for transfer passengers in a subway passage in Beijing and finds that the speeds differ significantly between males and females and between passengers walking alone and passengers in a company, with the walking speed generally following a log-normal distribution. A similar finding on the walking speeds following a log-normal distribution is reported in Zhu et al. (2017). Kasehyani et al. (2019) studies the walking times at different times of the day and finds that these differ, while other factors such as if passengers carry luggage also affects the walking speed.

Due to the varying walking speeds, the walking times at public transport stations are also not a constant factor of the distance walked. Daamen et al. (2006) studies passenger walking times for both boarding and alighting passengers at two stations in the Netherlands and specifically investigates which paths they use to and from the platform. By following passengers from the entrance to the station to the platform and vice versa, they find that passengers mainly choose the shortest path through the station. A similar methodology on following passengers to observe the walking times is used in Du et al. (2009), but with a focus on transferring passengers in Beijing. Significantly different walking times are found for passengers in the peak period and outside this period due to effects of crowding. The effect of crowding is also found to be significant in the study by Zhou et al. (2016) on walking speeds at different cross-sections of stations such as escalators, horizon passage and on the platform.

In recent years the focus has shifted from manual observations of walking times to estimations of the walking times based on smart card data. Smart card data is a valuable source for different types of analysis of passenger travel behaviour, such as travel time estimation, estimation of demand from origins to destinations and analysis of passenger route choice (Pelletier et al., 2011). The availability of the data is increasing in almost any major city and can help public transport agencies for better planning of the system and thereby for attracting more passengers to the system (Faroqi et al., 2018).

The vast majority of the studies using smart card data for estimation of walking times focus on the access and egress part of the trip from gate to platform and vice versa (Leurent and Xie, 2017; Li et al., 2020; Singh et al., 2020; Xie and Leurent, 2017), while only few studies focus on estimating the walking times at transfers (Jang, 2010; Sun et al., 2015; Wahaballa et al., 2018; Zhu et al., 2020), which are the times of interest in this paper. Jang (2010) use smart card data to detect transfer stations where the total transfer time is high with the aim of finding stations with bad coordination between bus and rail. Only aggregate results for all stations are provided and it is stated that it was not possible to split the transfer time in time used for walking and time used for waiting at the platform. Zhu et al. (2020) estimates the walking time of transferring passengers by finding the egress speed percentile of an individual passenger compared to other passengers. This percentile is used to find passengers' walking times at transfers by again comparing to the group of transferring passengers. The model is part of a complete approach for estimation of the total travel times from origin to destination. No validation of the transfer walking times are provided, other than fitted distributions of the walking times, which is a result of a fifth-degree polynomial estimation of the total travel time. Another study with a focus on transfer times, Wahaballa et al. (2018), studies the walking and waiting times at transfers between bus and rail using smart card data. The study proposes a stochastic frontier model, which aims at estimating the waiting time at transfers, while also considering the heterogeneity in walking times as these differ between passengers. The walking times can be observed from bus to the entry-gate, and these times are used directly as the walking time from rail to bus. A clear advantage of the smart card system used in the study, when considering walking times, is that the cards also are used for shopping and thereby these passengers are removed. No numbers are provided on the share of passengers shopping during the transfer, and hence it is difficult to tell how many observations can be removed due to this information. A final study using smart card data for estimation of transfer walking times is Sun et al. (2015), which use smart card data with tap-in and tap-out knowledge from bus to fare gantries at train stations to estimate the walking time based on a number of factors including card type and time of day. Using a linear regression model for each relation of bus stop and fare gantry they also include factors such as crowding and what is denoted collective pressure, i.e. whether the passengers walking speed is affected by the speed of passengers near them. The study does not account for whether passengers are doing activities during the study, as the time from tap-out to tap-in is used as the walking time proxy.

This information on whether a passenger is doing an activity during the transfer is not generally available in smart card systems and no studies investigating this have been found. However, Fujiyama and Cao (2016) has shed some light on this for terminal stations by studying the additional time spend at terminal stations in London before boarding the train. This can be observed, as passengers tap-in when entering the station and again near the platform. By assuming a general walking speed and a calibration for the individual paths made by the authors, they measure the additional time spend in the station. Interestingly, no correlation is found between the additional time spend and neither the total travel time or frequency of the line used. However, the additional time spent at the station is longer in the afternoon and evening compared to the morning.

3. Methodology

In this section the methodology is presented, along with preliminary requirements and data pre-processing needed prior to modelling. Figure 1 illustrates a transfer site, and the overall terminology for the proposed method. The goal is to estimate the walking time distributions for the different *path pairs* (4 shown in figure), without explicit knowledge of passengers true walking time nor knowledge on whether or not they performed an activity during their transfer.

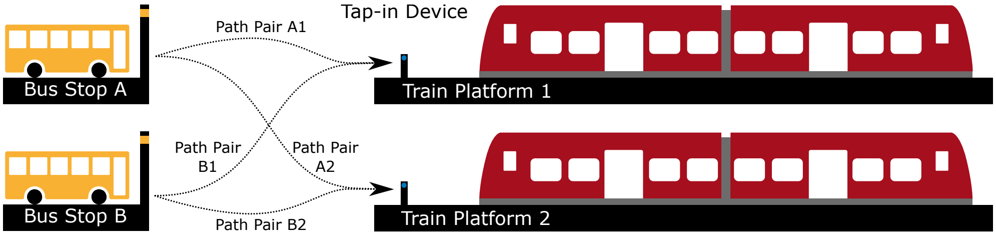


Figure 1: Overview of challenge and infrastructure setup.

We assume an AFC infrastructure, where tap-ins occur both when boarding a bus, and when entering a train platform. We assume the tap-in devices are located at platforms so it is possible to board a train immediately after tapping in. Passengers are assumed to tap-in when they enter the platform and not wait until departure of the train, which have also been showed to be true for the Danish smart card system (Ingvardson et al., 2018). Finally, we assume the passengers tap-out at their final destination.

3.1. Data Requirements and Pre-Processing

To apply the proposed method we need to prepare a data fusion between *AVL data* and *AFC data*. The following describes this fusion of data. We generally distinguish information belonging to the k'th stop of *bus trip* j (bus AVL dataset) and information belonging to the n'th trip leg of *passenger trip* i (AFC dataset).

We assume that the following information on bus AVL data is available or can be transformed to a similar structure. For each bus trip j we assume the availability of the following information:

- $\textit{Bus Ref}_j$: A unique reference to the vehicle that was observed running *bus trip* j.
- $\textit{Bus Stop Point Ref}_{j,k}$: A unique reference to k'th stop point for bus trip j which was observed arriving/departing.
- $\textit{Bus Arrival}_{j,k}$: Moment at which the vehicle was measured arriving to the k'th stop point of bus trip j.
- $\textit{Bus Departure}_{j,k}$: Moment at which the vehicle was measured departing from k'th stop point of bus trip j.

This information is standard output for most public transport AVL systems, and is included as part of the *GTFS-RT feed specification* (Google, 2020), although not all variables are considered mandatory.

From the AFC system we assume data is available or transformable to the following form:

- $\textit{Tap In}_{i,n}$: Moment at which the passenger tapped in for the n'th time on passenger trip i.
- $\textit{Bus Ref}_{i,n}$: A unique reference to the vehicle in which the $\textit{Tap In}_{i,n}$ occurred. For tap-ins conducted on train platforms $\textit{Bus Ref}_{i,n} = \emptyset$.
- $\textit{Stop Point Ref}_{i,n}$: A unique reference to bus stop point or train station platform this tap-in was conducted at.
- $\textit{Tap Out}_i$: The final tap out time for passenger trip i, i.e. at the passengers' destination.

We further assume the availability of a function $D(x, y)$ which measures the Euclidean distance between bus stop point x and train stations platform y. The euclidean distance function is solely used in the prepossessing of data to match Bus AVL and AFC data. The euclidean distance is used, since the network distance between bus stop and train station platform is not known.

The matching and data fusion between bus AVL and AFC data is a two-step process where we iterate AFC entries. First step is to match the passenger boarding to the bus AVL and secondly match the passenger alighting given the constraints of the boarding match. The match of the boarding is done by searching in bus AVL entries. For the n'th trip leg in passenger trip i we identify j and k by minimizing $|\textit{Tap In}_{i,n} - \textit{Bus Departure}_{j,k}|$ where $\textit{Bus Ref}_i = \textit{Bus Ref}_j$ and $\textit{Stop Point Ref}_{i,n} = \textit{Stop Point Ref}_{j,k}$. We denote the result of the boarding match:

$$\textit{Match Departure}_{i,n} \leftarrow (j, k)$$

We have now aligned information between bus AVL and AFC data for the boardings using tap-ins from AFC. To complete the second step we also want to match the alightings, and thus allowing the measurement of the observed walking time W^O. We need to identify the alighting stop k' prior to $\textit{Tap In}_{i,n}$ and do so by minimizing $D(\textit{Stop Point Ref}_{j,k'}, \textit{Stop Point Ref}_{i,n})$ where $j = \textit{Match Departure}^{j}_{i,n-1}$ and $k' > \textit{Match Departure}^{k}_{i,n-1}$. I.e. we search for the closest alighting stop on the matched *bus trip* j on the previous trip leg $(n-1)$ of *passenger trip* i. We constrain the search to only stops visited by the bus after the boarding stop. We denote the result of alighting stop match:

$$\textit{Match Arrival}_{i,n} \leftarrow (j, k')$$

The final result of the data pre-processing and matching process is a fused dataset for train tap ins (i.e. $\textit{Bus Ref}_{i,n} = \emptyset$), along with the matched bus alighting of the previous trip leg. Since we wish to estimate walking time for bus to train transfers we denote each combination of bus alighting stop point and train platform as a *path pair*. We split the data into separate data sets for each train station, and for each station data set we will consider the number of unique *path pairs* as $Q \in \mathbb{N}$ with $q \in \{1, \ldots, Q\}$. We denote the i'th observed walking time on *path pair* q as $W^O_{q,i}$.

3.2. Model

To model the behaviour of walking time during a transfer, we propose a hierarchical mixture model for each station with transfers of bus stop to train stations. Each station will have Q path pairs, where the observed variable is the walking time $\boldsymbol{W}^O_q \in \mathbb{R}^{N_q}$ of $N_q \in \mathbb{N}$ trips along the q'th path pair. The observed walking time is assumed to originate from two types of unobserved walking behaviours Ω^k_q for $k \in \{D, A\}$: (i) passengers walking directly, and (ii) passengers doing an activity during the transfer with the following definitions:

Definition 1. *The direct walking behaviour, Ω^D, describes a transfer done by a passenger who walks directly from a bus stop to a train platform with a normal walking speed, thereby having a direct walking time W^D.*

Definition 2. *The activity walking behaviour, Ω^A, describes a transfer, where an activity affects the walking time, such as shopping, buying coffee, etc, thereby having an activity-based walking time W^A.*

Definition 3. *For each observed walking time $W^O_{q,i}$ there is an individual unobserved behavior $Z_{q,i} \in \{D, A\}$.*

Definition 4. *The share of passengers walking directly on the q'th path is $\lambda_q \in [0, 1]$.*

The posterior distribution of the walking time behaviors and the share of passengers walking directly given observed walking time, can be written using Bayes' rule as

$$P(\boldsymbol{\Omega}, \boldsymbol{\lambda} | \boldsymbol{W}^O) \propto P(\boldsymbol{W}^O | \boldsymbol{\Omega}, \boldsymbol{\lambda}) P(\boldsymbol{\Omega}, \boldsymbol{\lambda}). \tag{1}$$

To obtain the final model the following assumptions are made relating to the share of passengers walking directly, the walking time and path pairs:

Assumption 1 (Origin of walking time)**.** *It is assumed that the i'th walking time, $W^O_{q,i}$, originates from either walking directly ($Z_i = D$) or activity-based walking ($Z_i = A$). When the walking behaviour is known, the walking time only depends on the given behaviour e.g. the conditional distribution of walking time $P(W^O_{q,i} | \Omega_q, Z_{q,i} = D)$ is independent of the activity-based behaviour Ω^A and $P(W^O_{q,i} | \Omega_q, Z_{q,i} = A)$ is independent of the direct walking behaviour Ω^D, which gives*

$$P(W^O_{q,i} | \Omega_q, Z_{q,i} = D) = P(W^O_{q,i} | \Omega^D_q) \quad \textit{and} \quad P(W^O_{q,i} | \Omega_q, Z_{q,i} = A) = P(W^O_{q,i} | \Omega^A_q). \tag{2}$$

From definition 1 and 2 we also get the walking time for the two behaviours by

$$P(W^D_{q,i}) = P(W^O_{q,i} | \Omega^D_q) \quad \textit{and} \quad P(W^A_{q,i}) = P(W^O_{q,i} | \Omega^A_q). \tag{3}$$

Assumption 2 (Time invariant of walking behavior shares)**.** *The share of passengers walking directly $\lambda_q \in [0, 1]$ is assumed to be constant over time, and since the walking behaviour can only be direct or activity-based (assumption 1), the share of passengers doing an activity is given by $(1 - \lambda_q)$. Given this assumption, when we know the share of passengers walking directly, then the probability of the i'th passenger walking directly and the probability of the i'th passenger doing an activity, e.g. the conditional probability for a passenger walking behaviour conditional on the share of passengers walking directly, is*

$$P(Z_i = D | \lambda_q) = \lambda_q \quad \textit{and} \quad P(Z_i = A | \lambda_q) = (1 - \lambda_q). \tag{4}$$

Assumption 3 (Independent path pairs and trips walking time)**.** *It is assumed that the walking time of trip i is independent of the walking time of all other trips and that all path pairs are independent of all other path pairs, such that*

$$P(\boldsymbol{W}^O | \boldsymbol{\Omega}, \boldsymbol{\lambda}) = \prod_{q=1}^{Q} \left[\prod_{i=1}^{N_q} P(W^O_{q,i} | \Omega_q, \lambda_q) \right]. \tag{5}$$

The assumption of independent walking time will not be valid for stations with congestion due to crowding. However, crowding at transfers is not a major issue at the stations included in this analysis.

Assumption 4 (Independence between walking types)**.** *It is assumed that the share of passengers walking directly is independent of the direct and activity-based walking behaviour and that the behaviours only depend on its own given behaviour such that*

$$P(\boldsymbol{\Omega}, \boldsymbol{\lambda}) = P(\boldsymbol{\Omega}^D, \boldsymbol{\Omega}^A, \boldsymbol{\lambda}) = P(\boldsymbol{\Omega}^D) P(\boldsymbol{\Omega}^A) P(\boldsymbol{\lambda}). \tag{6}$$

It could be argued that the activity-based walking behaviour depends on the direct walking time since an activity will add to the total time of the transfer. This may result in a bias in the activity-based walking time. The walking time distributions for the two behaviours should be constrained to mitigate this relation between the expected walking times $\mathbb{E}(W^D)$ and $\mathbb{E}(W^A)$.

Using equation 1 in combination with the assumption 3 and 4 relating to the path pairs and walking time, we can derive

$$P(\boldsymbol{W}^O|\boldsymbol{\Omega}, \boldsymbol{Z}) = \prod_{q=1}^{Q} \left[\prod_{i=1}^{N_q} P(W_{q,i}^O|\Omega_q, \lambda_q) \right] P(\boldsymbol{\Omega}^D)P(\boldsymbol{\Omega}^A)P(\boldsymbol{\lambda}). \tag{7}$$

The likelihood $P(W_{q,i}^O|\Omega_q, \lambda_q)$ can be rewritten by first marginalizing over possible behaviours, direct walk D and activity walk A, and then applying assumption 1 and 2, such that we get

$$\begin{aligned} P(W_i^O|\Omega_q, \lambda_q) &= P(W_{q,i}^O, Z_i = D|\Omega_q, \lambda_q) + P(W_{q,i}^O, Z_i = A|\Omega_q, \lambda_q) & (8) \\ &= P(Z_i = D|\lambda_q)P(W_i^O|Z_i = D, \Omega_q) + P(Z_i = A|\lambda_q)P(W_{q,i}^O|Z_i = A, \Omega_q) & (9) \\ &= \lambda_q P(W_{q,i}^O|\Omega_q^D) + (1 - \lambda_q)P(W_{q,i}^O|\Omega_q^A). & (10) \end{aligned}$$

Then inserting equation 10 into equation 7 we get the the final equation

$$P(\boldsymbol{\Omega}, \boldsymbol{\lambda}|\boldsymbol{W}^O) \propto \prod_{q=1}^{Q} \left[\prod_{i=1}^{N_q} \left[\lambda_q P(W_{q,i}^O|\Omega_q^D) + (1 - \lambda_q)P(W_{q,i}^O|\Omega_q^A)\right] \right] P(\boldsymbol{\Omega}^D)P(\boldsymbol{\Omega}^A)P(\boldsymbol{\lambda}). \tag{11}$$

Figure 2 shows a graphical representation of equation 11, which have the form of a hierarchical mixture model, with the likelihood being $P(W_{q,i}^O|\Omega_q^A)$ and $P(W_{q,i}^O|\Omega_q^D)$ as mixture components, the mixture weights λ, and the priors $P(\boldsymbol{\Omega}^D)$, $P(\boldsymbol{\Omega}^A)$ and $P(\boldsymbol{\lambda})$.

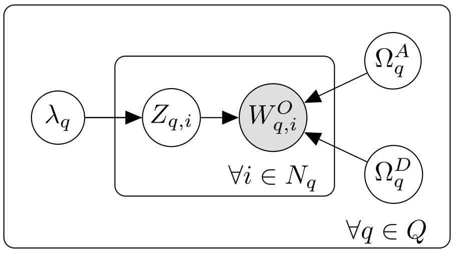


Figure 2: The model represented as a probabilistic graphical model.

The mixture components $P(W_{q,i}^O|\Omega_q^A)$ and $P(W_{q,i}^O|\Omega_q^D)$, expressing the direct and activity-based walking time (assumption 1), are both assumed to Beta-distributed, where each behaviour Ω contains the Beta's shape parameters (α, β).

$$\begin{aligned} P(W_{q,i}^D) &= P(W_{q,i}^O|\Omega_q^B) \sim Beta(\alpha_q^D, \beta_q^D) \quad \text{and} \\ P(W_{q,i}^A) &= P(W_{q,i}^O|\Omega_q^A) \sim Beta(\alpha_q^A, \beta_q^A). \end{aligned}$$

The posterior distribution is the combination of the likelihood and the prior. The prior encode domain knowledge apriori into the model's parameters. The degree of domain knowledge encoded into the prior determines how uncertain the model is about the true values before seeing any data and how much data is needed to be seen before likelihood dominates the posterior distribution. A common way to categories priors are into non-informative, weakly informative and informative priors (Sarma and Kay, 2020). Suppose the domain knowledge tells that some values are unrealistic. In that case, an informative prior can be used to concentrate the probability mass on a small range of the value space, thereby making the unrealistic values unlikely. This means that strong evidence from the likelihood is needed before the posterior distribution is dominated by likelihood. If little is known about the outcome values, a non-informative prior can spread the probability mass out, making all values equally likely. A weakly informative prior can be used between these two cases, where the likelihood can easily dominate the posterior distribution if the data is adequately informative.

For the proposed model, the prior on the share of passengers transferring directly $\lambda_q \sim Beta(4, 2)$ is weakly informed since it is reasonable to expect that most passengers walk directly. The priors on the walking behaviours Ω are both constrained non-informative priors. It is assumed, for the direct walking behaviour $\Omega^D \in (\alpha^D, \beta^D)$, that the expected direct walking time $\mathbb{E}(W^D)$ cannot exceed the 15 minutes, which is half of the total maximum allowed transfer time of 30 minutes in the danish AFC system. This assumption is obtained by the constraint $\alpha_q^D \leq \beta_q^D$. From definition 2 the activity-based behaviour Ω^D should have a high walking time with a large variation. To model this behaviour, we assume the expected activity-based walking time $\mathbb{E}(W^A)$ to be between 12 and 18 minutes by constraining $\alpha^A \in [2, 3]$ and $\beta^A \in [2, 3]$.

Finally, to estimate each walking behaviour, we use the posterior predictive distribution for each walking behaviour, such that

$$P(\widehat{\boldsymbol{W}}|\boldsymbol{W}^O) = \iiint P(\widehat{\boldsymbol{W}}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\lambda})P(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\lambda}|\boldsymbol{W}^O)\, d\boldsymbol{\alpha}\, d\boldsymbol{\beta}\, d\boldsymbol{\lambda}, \tag{12}$$

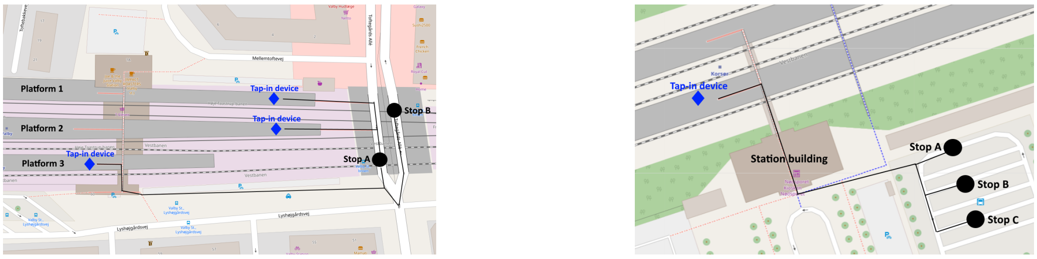
$$P(\widehat{\boldsymbol{W}}^A|\boldsymbol{W}^O) = \iint P(\widehat{\boldsymbol{W}}^A|\boldsymbol{\alpha}^A, \boldsymbol{\beta}^A, \boldsymbol{Z} = A)P(\boldsymbol{\alpha}^A, \boldsymbol{\beta}^A, |\boldsymbol{W}^O)\, d\boldsymbol{\alpha}^A\, d\boldsymbol{\beta}^A \quad \text{and} \tag{13}$$

$$P(\widehat{\boldsymbol{W}}^D|\boldsymbol{W}^O) = \iint P(\widehat{\boldsymbol{W}}^D|\boldsymbol{\alpha}^D, \boldsymbol{\beta}^D, \boldsymbol{Z} = D)P(\boldsymbol{\alpha}^D, \boldsymbol{\beta}^D, |\boldsymbol{W}^O)\, d\boldsymbol{\alpha}^D\, d\boldsymbol{\beta}^D. \tag{14}$$

4. Case study

Our case study is conducted for the entire Eastern Denmark for November 2019. We include most train stations serviced by the national rail service provider, metro stations and some local train stations. Figure 5 shows a map of the included stations. The model was estimated on 129 stations with a total of 1,009 path pairs. Only path pairs with 100 or more observations during November were estimated, as these pairs then have an average of at least three transferring passengers pr. day. The final dataset consists of 542,713 observations, i.e. unique transfers. Each station was estimated separately by the probabilistic language STAN using NUTS sampling with four chains, each with 3,000 iterations, and a warm-up period of 2,000 iterations. Since it is not feasible to present all the results in detail, two stations have been selected for detailed analysis of the results and verification of the model assumptions.

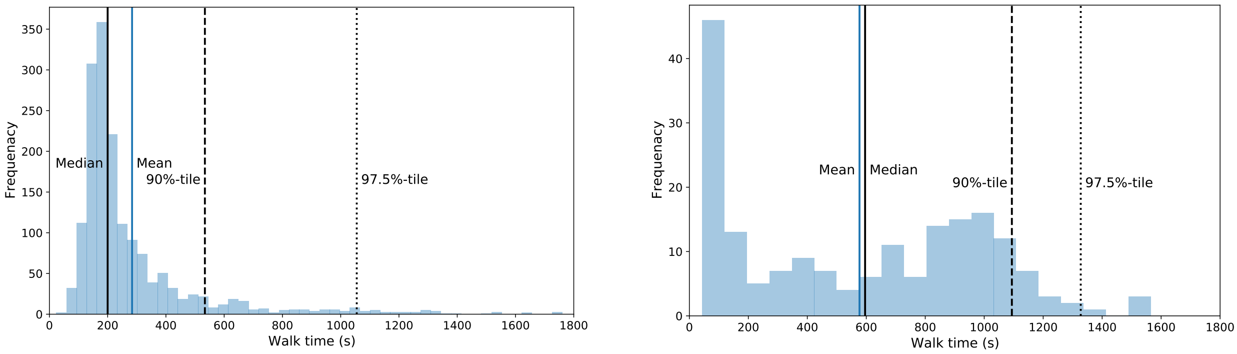
To illustrate and analyse the model estimations in more detail, the stations at Valby (case 1) and Korsør (case 2) will be used as examples. As a larger transfer station in the Copenhagen area, Valby Station has an expected distribution of the walking times as shown in Figure 4a, where most passengers have a relative low walking time. The layout of the station is presented in Figure 3a, where the path pairs selected for the analysis are also presented. In contrast to Valby, Korsør is a small rural station with an abnormal observed walking time distribution with two peaks shown in Figure 4b. The first peak has the expected location of a relative low walking time, where second peak is located above the median of 10 minutes. The station layout of Korsør station is shown in Figure 3b. The station building includes a waiting hall and a convenience store.



(a) Overview of Valby Station

(b) Overview of Korsør Station

Figure 3: Overview of station layouts for selected stations. Background source: OpenStreetMap



(a) Observed walking times (from alighting the bus until tap-in at platform) from AFC data at Valby Station (urban station on Zealand, Denmark), specifically path pair connecting bus stop B to platform 3 (V-B3) which has an approximate network walking distance of 200 meters.

(b) Observed walking times (from alighting the bus until tap-in at platform) from AFC data at Korsør Station (rural station on Zealand, Denmark), specifically path pair connecting bus stop B to the platform (K-B) which has an approximate network walking distance of 150 meters.

Figure 4: Histograms of the raw walking time observations for two stations (selected path-pairs)

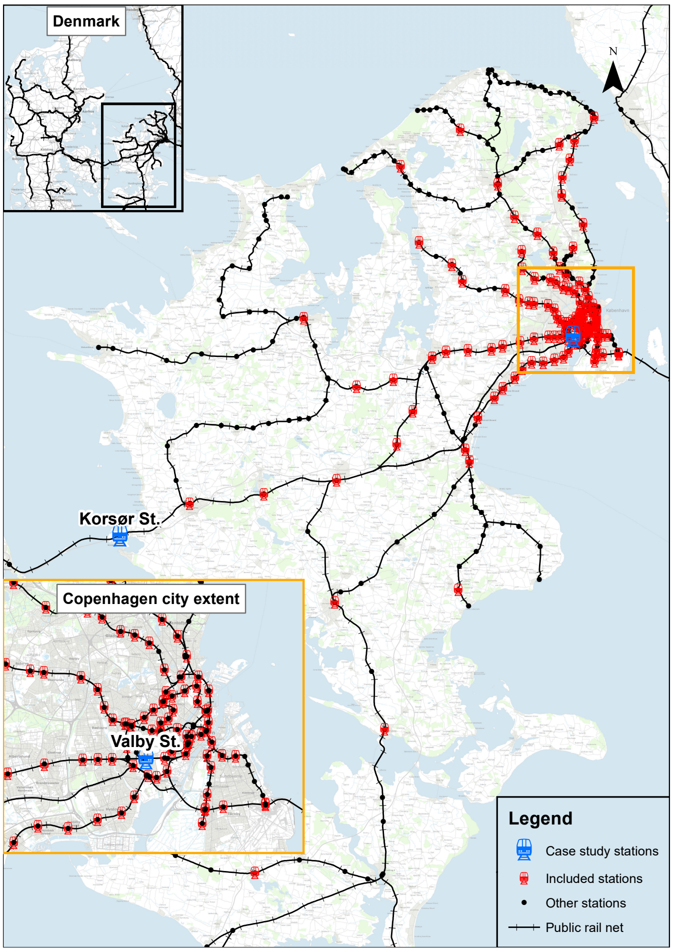


Figure 5: Overview of included stations in the analysis. Background map source: GeoDanmark-data (2020)

4.1. Case station 1: Valby

Valby has 32 different path pairs, where we have selected the results from six path pairs, which are combinations of the two bus stops and three platforms shown in Figure 3a. The six path pairs include a total of 11,875 observations, which is a subset of the 19,439 total observations at Valby Station. The four path pairs V-A1, V-A2, V-B1 and V-B2 are transfers to platforms used by suburban train services, whereas V-A3 and V-B3 are transfers to regional trains. Table 1 shows that the path pair with the largest distance V-A3 and V-B3 have the highest mean observed walking time with respectively 4 and 5 minutes. From stop B the passengers walking have to cross a pedestrian crossing to get to the different platforms, which results in a mean difference between stop A and B of 50 seconds on average.

The observed walking times are compared to the scheduled walking time, which is used in travel planners (Rejseplanen (Danish Travel Planner), 2020) and for coordinating buses and trains. This shows that at least 4% of the passengers transferring to platform 1 and 2 are not able to make the scheduled transfer time, where in the case of V-A3 and V-B3 there are respectively 27% and 35%. If the raw walking time was to be used as an indicator for the direct walking time, the scheduled walking time for both stops to platform 3 should be increased to accommodate the higher walking times.

		Observed walking time					Scheduled Walking time	
Path Pair	N	Mean	Std	2.5%-tile	Median	97.5%-tile	Value	Above
V-A1	2206	90.61	121.15	40.00	65.00	426.62	240	4.26%
V-A2	523	103.18	180.29	35.05	60.00	541.90	240	6.69%
V-A3	3460	244.72	235.25	82.00	149.00	981.93	240	27.57%
V-B1	2878	142.86	107.70	68.00	119.00	394.00	240	5.77%
V-B2	1153	159.08	139.79	77.00	126.00	579.00	240	7.37%
V-B3	1655	283.68	236.92	100.00	199.00	1054.95	240	35.59%

Table 1: Observed walking time and Schedule walking time of path pairs at Valby.

Table 2 presents the results of the model for both the share of passengers walking directly, the direct walking time $\widehat{W}^D$, walking time for passengers with activity $\widehat{W}^A$ and the predictive posterior distribution $\widehat{W}$ from each stop to the three platforms. If we compare the direct walking time $\widehat{W}^D$ to the scheduled walking time, there is larger share of the passengers that are able to make the transfer compared to the observed walking time. All transfers for direct walking passengers to platform 1 and 2 have less than 1% of the density above the scheduled walking time, where V-A3 has 1.35% and V-B3 has 24.85% above. For path pair V-A1 and V-A2 it is possible to reduce the scheduled walking time to 2 minutes and still have less than 1% of the density above the scheduled walking time.

Continuing to the fit of the model, we see that the direct walking time $\widehat{W}^D$ aligns with the differences between path pairs described for the observed walking time. The highest walking times from both bus stops are found for passengers walking to platform 3, and the model estimates that it takes on average 1 minute longer for passengers to walk from stop B than stop A. A visual inspection of the model estimations in Figure 6A and the predictive posterior walking $\widehat{W}$ shows that all path pairs have a peak at the same position as the observed walking time followed with a long tail. The peak originates from the direct walking time distribution shown in Figure 6B, where the long tail originates from the activity. The figure shows that the density of the activity distribution ranges over the direct walking time distribution, which results in underestimating the direct walking share λ giving the smaller peak of the predictive posterior walking compared to the observed walking time. We can examine the predictive posterior walking time $\widehat{W}$ closer in Table 2 by comparing it with the observed walking time from Table 1. Subtracting the median observed walking time of each path pair with the corresponding predictive posterior walking time $\widehat{W}$, the predictive posterior walking time is on average 4.7 seconds higher than the median observed walking time. For the upper percentiles, the estimation is notable above the observed walking time, supporting the visual inspection of the underestimation of the direct walking share. The model estimates a share of passengers walking directly λ ranging from 73% to 95%, where the two lowest λ values (73% and 80%) are estimated for path pairs to

Parameters	ID	Mean	Sd	2.5%-tile	Median	97.5%-tile	ess	$\hat{R}$	Above scheduled time
λ	V-A1	0.93	0.01	0.92	0.93	0.94	5958	1.0	-
	V-A2	0.89	0.01	0.86	0.89	0.92	7274	1.0	-
	V-A3	0.73	0.01	0.71	0.73	0.74	4574	1.0	-
	V-B1	0.95	0.00	0.94	0.95	0.96	6428	1.0	-
	V-B2	0.93	0.01	0.92	0.93	0.95	6691	1.0	-
	V-B3	0.80	0.01	0.78	0.80	0.83	5009	1.0	-
$\widehat{W}^A$	V-A1	724.77	365.30	113.18	695.17	1462.60	3970	1.0	-
	V-A2	728.12	361.87	129.06	703.23	1460.16	4000	1.0	-
	V-A3	724.14	358.89	121.91	694.96	1457.00	3650	1.0	-
	V-B1	725.68	362.40	120.31	706.11	1461.97	3895	1.0	-
	V-B2	735.94	361.96	128.52	712.20	1462.73	4025	1.0	-
	V-B3	728.23	357.04	122.51	711.24	1455.79	3964	1.0	-
$\widehat{W}^D$	V-A1	66.64	16.63	37.93	65.20	102.55	4089	1.0	0.00%
	V-A2	61.58	16.54	33.39	60.14	97.70	4202	1.0	0.00%
	V-A3	142.63	39.08	76.35	139.00	228.43	3643	1.0	1.35%
	V-B1	124.82	36.65	63.08	121.48	205.15	3700	1.0	0.45%
	V-B2	130.05	37.27	65.68	126.91	208.48	4083	1.0	0.43%
	V-B3	198.52	65.34	89.80	192.48	343.95	4008	1.0	24.85%
$\widehat{W}$	V-A1	112.99	187.74	38.51	66.84	804.97	3834	1.0	10.72%
	V-A2	139.13	245.37	33.84	62.72	1036.08	3995	1.0	6.82%
	V-A3	301.25	324.48	78.56	155.33	1246.30	3903	1.0	25.85%
	V-B1	159.09	169.34	63.13	124.04	763.39	4119	1.0	5.65%
	V-B2	169.55	177.39	66.06	129.79	827.76	4015	1.0	6.42%
	V-B3	294.88	263.69	90.96	207.85	1162.98	3896	1.0	37.05%

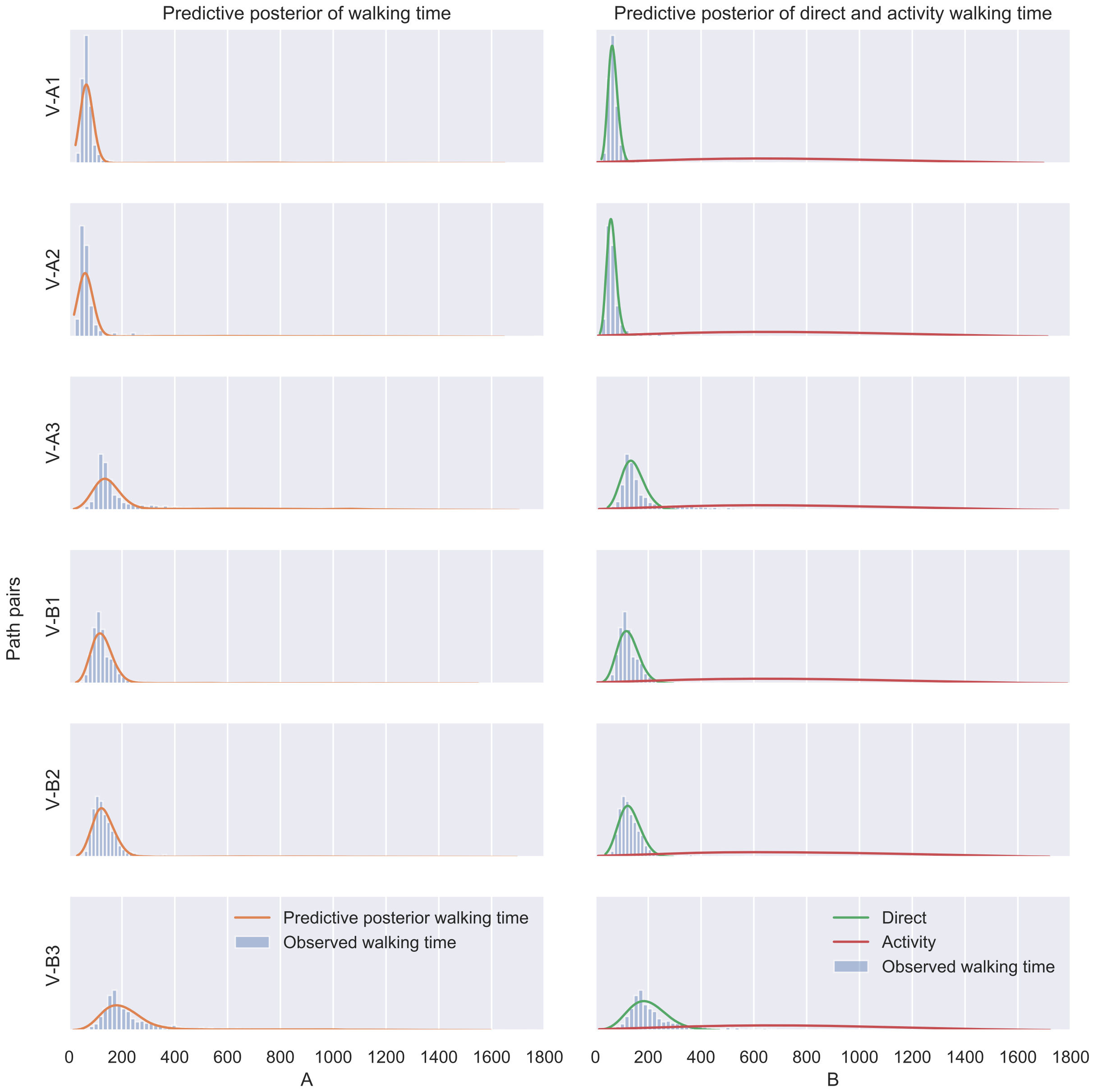
Table 2: Valby - Posterior means and statistics in seconds.

the regional train services at platform 3. Compared to the suburban rail services on platform 1 and 2, the headway is larger for the regional, making it easier for passengers to do an activity without missing their train.

4.2. Case station 2: Korsør

Korsør has three path pairs shown in Figure 3b, which are three different bus stops to the same platform at the station. As shown in Table 3 the path pair K-C has the lowest mean observed walking time of 3.7 minutes in combination with the highest schedule walking time of 4 minutes compared to the two others path pairs schedule walking time of 3 minutes. With a lower scheduled walking time, it would be expected, that the observed mean walking time would be smaller for the path pairs K-A and K-B, but we can see from the Table 3 that the mean walking time is nearly double for both. Comparing the scheduled walking time to the observed, we see that path pair K-C has 30% of the observed walking time above the scheduled walking time, while path pairs K-A and K-B have respectively 50% and 70% above. Using the raw walking time as an indicator for the needed walking time, would thus increase the scheduled walking time significantly.

The model estimates a low degree of the transfer passengers walking directly from the bus to the station, where the mean share of passengers walking directly ranges from 26% to 63%. The highest activity share is the path pair K-B, which was suspected of having an abnormal transfer pattern. If we look at the posterior predictive walking time $\widehat{W}$ of the path pair K-B in Figure 7A, we see that a large part of the density is spread in the tail. At the same time, we see a significant number of the observed walking time samples are located here, thus supporting the high degree of activity. A comparison between the distribution of $\widehat{W}$ in Table 4 and the observed walking time in Table 3 shows a reasonable match between the two. The fit does



(A) The predictive posterior walking time distribution is generated from the weighting of the direct walking share of activity and direct walking time. (B) The distribution of activity and direct walking time distribution without the direct walking share.

Figure 6: Valby station - Predictive posterior of walking time compared to observed walking time.

not seem as good as for the other case station, since the lower percentiles underestimates and the upper percentiles overestimates values. Looking at the predictive posterior of the directly and activity walking time we see separated peaks for the two distributions, but there are, as with the estimation for the other case station, areas where the density of the activity and directly walking time overlaps. This could possibly affect the models ability to separate the two distributions.

If we compare the scheduled walking time to the direct walking time $\widehat{W}^D$ distribution, less than 1% of the density is above the scheduled walking time for the three path pairs K-A, K-B and K-C, which indicates

				Observed walking time			Scheduled walking time	
Path Pair	N	Mean	Std	2.5%-tile	Median	97.5%-tile	Value	Above
K-A	130	427.32	449.57	59.22	199.00	1443.10	180	50.77%
K-B	187	577.06	416.16	56.00	596.00	1327.40	180	69.52%
K-C	386	227.41	260.37	41.87	94.50	960.12	240	29.53%

Table 3: Korsør - Observed walking time and Schedule walking time of path pairs.

Parameters	ID	Mean	Sd	2.5%-tile	Median	97.5%-tile	ess	$\hat{R}$	Above scheduled time
λ	K-A	0.63	0.03	0.57	0.63	0.68	2779	1.0	-
	K-B	0.26	0.04	0.19	0.26	0.34	2059	1.0	-
	K-C	0.52	0.06	0.41	0.52	0.63	2536	1.0	-
$\widehat{W}^A$	K-A	730.26	359.86	119.92	709.46	1471.66	4042	1.0	-
	K-B	742.19	359.35	142.14	726.35	1460.54	3933	1.0	-
	K-C	755.01	366.21	143.12	728.99	1490.58	3933	1.0	-
$\widehat{W}^D$	K-A	78.88	28.54	32.26	75.27	141.04	3906	1.0	0.35 %
	K-B	75.94	21.41	42.27	73.06	124.90	3743	1.0	0.18 %
	K-C	100.36	42.17	37.95	93.46	200.68	3799	1.0	0.88 %
$\widehat{W}$	K-A	314.31	381.38	35.96	99.72	1307.28	4138	1.0	34.35 %
	K-B	577.97	426.36	51.14	549.11	1428.90	3900	1.0	71.90 %
	K-C	408.92	411.18	43.10	167.40	1378.21	4121	1.0	44.35 %

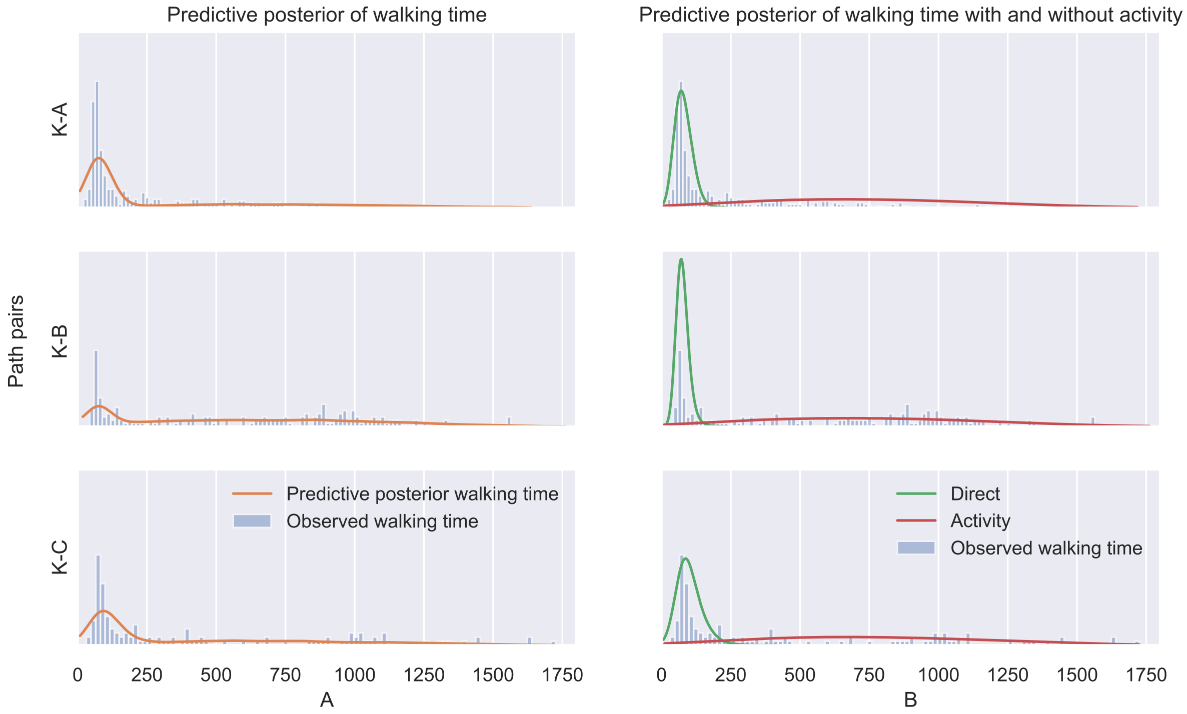
Table 4: Korsør - Posterior means and statistics in seconds.

that the scheduled walking times are reasonable. This indicates that the long duration of the observed walking times, is due to a high degree of activity at the station or lack of coordination between the bus and train schedule.

4.3. All stations

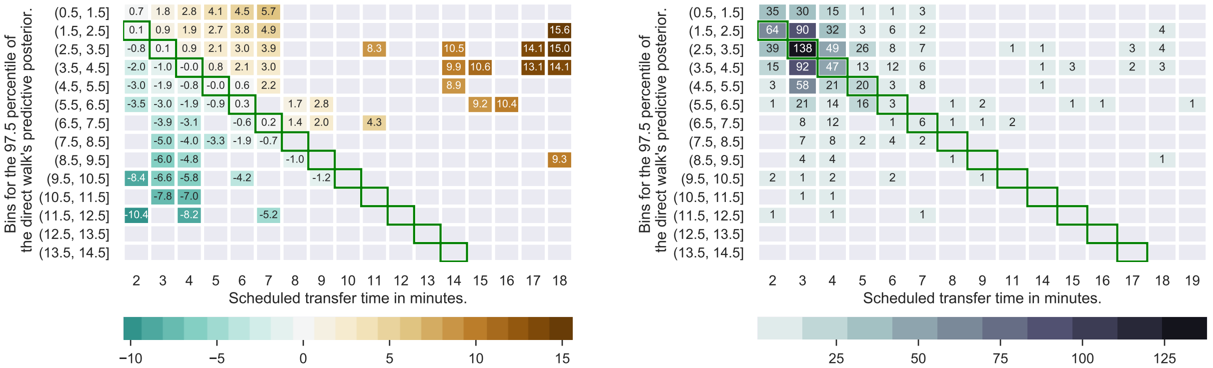
Finally, results for all 129 stations are collectively analysed and presented. The analysis focus on comparing the walking time as estimated by our method with the scheduled transfer time for each of the more than 1000 path pairs between bus stop points and train platform validators.

The result of the analysis is visually illustrated in Figure 8, showing the average difference between scheduled walking time and the 97.5$^{\text{th}}$ percentile of the estimated transfer walking time distribution (8a), respectively the number of path pairs (8b). The majority of the path pairs' scheduled time are between two and five minutes (8b), which is also reflected in the large number of path pairs being estimated in this interval. However, there are also several path pairs with a scheduled walking time above 14 minutes which are estimated to take 2-5 minutes, and also some path pairs estimated to take above 8.5 minutes with a scheduled walking time at 2-5 minutes. The path pairs with a scheduled walking time above 14 minutes are all scheduled with a too high walking time compared with the estimated walking time with an average difference ranging from 8.9 to 15.6 minutes. The majority of path pairs with an estimated walking time above 8.5 minutes have a too low scheduled walking time. The estimated walking time for these pairs have an average difference below -4.2 minutes of the scheduled walking time. The figures in Figure 8 makes it easy for the transit agencies to identify the path pairs, which needs further investigation to ensure that most passengers can catch their train when transferring from bus to train. In addition, by using the Bayesian framework, the transport agencies can use the inferred direct walking time distribution to set the level of passengers who should be able to catch their train if walking directly.



(A) The predictive posterior walking time distribution is generated from the weighting of the direct walking share of activity and direct walking time. (B) The distribution of activity and direct walking time distribution without the direct walking share.

Figure 7: Korsør station - Predictive posterior of walking time compared to observed walking time.



(a) Average difference between scheduled and estimated walking time.

(b) The count of the number of path pairs for each comparison of estimated and scheduled walking time.

Figure 8: Comparison of scheduled and estimated walking time on the path pairs of the 129 stations. The green squares indicate where the scheduled walking time is within 30 seconds of the 97.5th percentile of the estimated walking time.

5. Discussion

The validation of the proposed method is indeed difficult. As described in Section 3.1 we do not assume ground truth about whether passengers transferred directly is available, nor do we assume availability of their true walking time or choice of path.

As a consequence of the desire for a general and large scale applicable solution, manual validation in the form of accompanying or somehow recording passengers during their transfers in order to determine their true walking time and possible time used for activities were deemed infeasible. Such an approach would be both error-prone due to the human factor, and very time-consuming for collection of a representative sample. It can also be argued, that people might not recollect doing activities during transfers as for example used in Mosallanejad et al. (2018) for splitting trip chains into separate trips. On top of this, passengers also have difficulties in reporting reasonable walking times in surveys (Anderson, 2013). Therefore validation with classic surveys and interviews are considered insufficient and impractical.

To overcome this challenge we suggest two generalizable verification approaches that are applicable at scale: (i) *Verification using number of feasible trains*; and (ii) *Verification using shop availability data.* In the following sections we detail the two verification approaches. We recognize that the verification can be further improved for concrete cases, depending on the data available.

5.1. Verification of model results using number feasible trains

Verification using train assignment requires access to train AVL data similar to the bus AVL data described in Section 3.1. We only consider passengers who finished their journey after riding the train, i.e. $\textit{Tap In}_{i,n}$ is the train tap in on the last trip leg, n, for passenger trip i. We assign each passenger a set of feasible trains which runs directly to the destination station based on $\textit{Tap In}_{i,n}$ and $\textit{Tap Out}_i$. We likewise assign each passenger a set of feasible trains based on $\textit{Bus Arrival}_{j,k'}$ and $\textit{Tap Out}_i$, where $(j, k') = \textit{Match Arrival}_{i,n-1}$. The latter one corresponds to feasible trains given the passenger had absolutely no walking time at all.

Table 5 shows the number of observations decomposed by the number of feasible trains based on the two approaches for train assignment cf. above for passengers at Valby station.

	Using $\textit{Tap In}_{i,n}$				
Using $\textit{Bus Arrival}_{j,k'}$	1	2	3	4	**Total**
1	9,287				**9,287**
2	2,421	90			**3,323**
3	270	351	117		**738**
4	79	116	90	111	**396**
Total	**12,057**	**1,369**	**207**	**111**	**13,744**

Table 5: Decomposition of feasible trains for Valby station by approach.

The table indicates that around 85% of the passengers have only one feasible train given their $\textit{Tap In}_{i,n}$ time and the final $\textit{Tap Out}_i$. Most of these passengers have also only one feasible train given the $\textit{Bus Arrival}_{j,k'}$ time. However, there are also a considerable number of passengers who have a difference in the number of feasible trains given the two criteria. Some of these passengers, especially those who have a large difference on the number of feasible trains given the two criteria, are possibly more likely to have had an activity during the transfer, as they did not board some possible trains they could have caught if they walked fast to the platform. To test how the model predicts passengers within these groups, the observations can be combined with the prediction of the model. 4,000 samples of the set of parameters in the model are used to categorise passengers into three groups:

- Directly - All sampled sets of parameters assigned the highest probability of the observation belonging to the directly walking distribution.

- Activity - All sampled sets of parameters assigned the highest probability of the observation belonging to the activity walking distribution.

- Mixed - The observation was not consistently assigned to one of the groups.

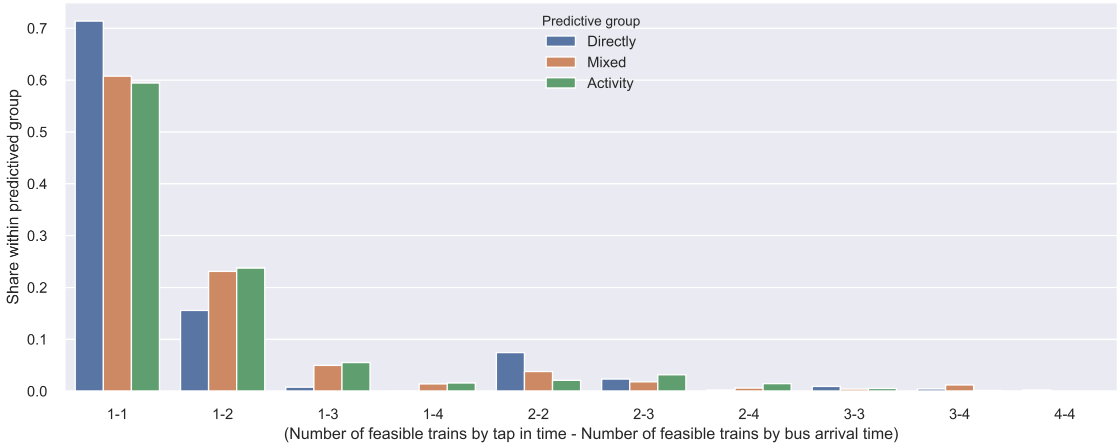


Figure 9: Distribution of passenger groups predicted to respectively walk, do an activity, or not uniquely identified, across combinations of number of feasible trains (tap in time vs. bus arrival time)

Figure 9 presents the share of passengers within each of the predicted groups belonging to the combination of each count of feasible trains. There is a noticeable difference between the distribution of passengers in the respective groups across the different combinations. For the group predicted to walk directly, around 70% of these have only one feasible train given both their tap in time and the arrival time of the bus. The shares for the group predicted to have an activity during the transfer is lower for this combination, and instead higher for the combination with two feasible trains given the bus arrival time and only one feasible train given the tap in time. The result that almost no passengers predicted to walk directly is placed in the group with three feasible trains given the bus arrival time and only one feasible train given the tap in time is reassuring, as this cluster indicates that the passenger could have possibly reached at least one train prior to the one boarded.

At Korsør station the dataset consists of 490 passengers who tapped out at the end of the train leg. Only 10 of these passengers had more than one feasible train given the bus arrival time, and hence the long observed walking times found in Section 4.2 stems from passengers who spend time at the station building instead of walking directly to the platform. The long walking times are thus an effect of the long transfer times, due to the lack of coordination between busses and trains.

5.2. Verification using shop availability data

One of the main assumptions for passengers not walking directly during the transfer is shopping activities. In order to support this assumption and provide a weak, but scalable verification of the proposed method, the share of activity transfers $(1 - \lambda_q)$ is correlated with shop availability. Since a unique value of λ_q per path pair q is obtained, we also use this granularity for shop availability.

Data is extracted from Open Street Map (OpenStreetMap contributors, 2018) using a buffer zone around the crow flies distance of path pair q as illustrated by Figure 10. The size of the buffer zone has been fixed to 500m in this experiment. We search this buffer zone using the Open Street Maps tag features, specifically nodes containing the tag `shop`.

We denote the number of shops in the buffer zone formed from path pair q as $\mathit{Shop\ Availability}_q$, and investigate the correlation between $1 - \lambda_q$ and $\log(\mathit{Shop\ Availability}_q)$. We apply the logarithm based on an expectation that the marginal effect of extra shops will eventually have a limited effect on how many passengers will take advantage of the availability.

Figure 11 shows the relation between $1 - \lambda_q$ and $\log(\mathit{Shop\ Availability}_q)$. We see a positive correlation between the two variables. The result supports some relationship between the estimated activity share for each path pair, and the shop availability along the path pair. Although the relationship is clearly not linear

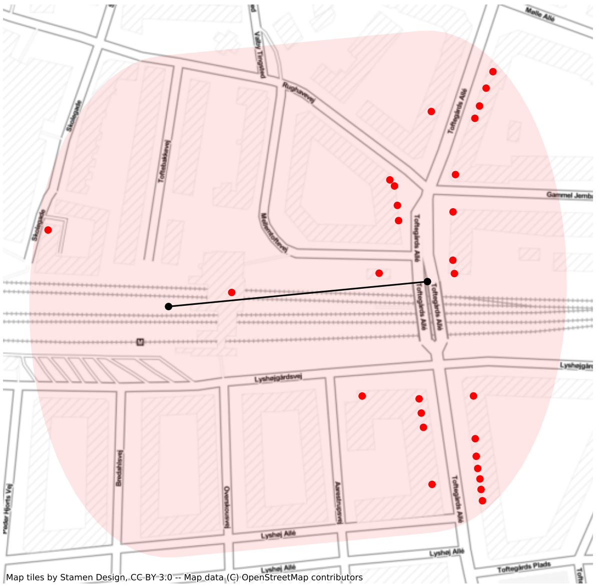


Figure 10: Example of shop availability buffer zone for Valby Station. Crow flies distance of path pair (black), Buffer zone (transparent red), Shops (red). Background source: OpenStreetMap

($R^2 = 0.25$), given that a high availability of shops does not guarantee a high share of passengers with activities. On the other hand, in all cases where the presented method has estimated high activity transfer share, we find a high availability of shops.

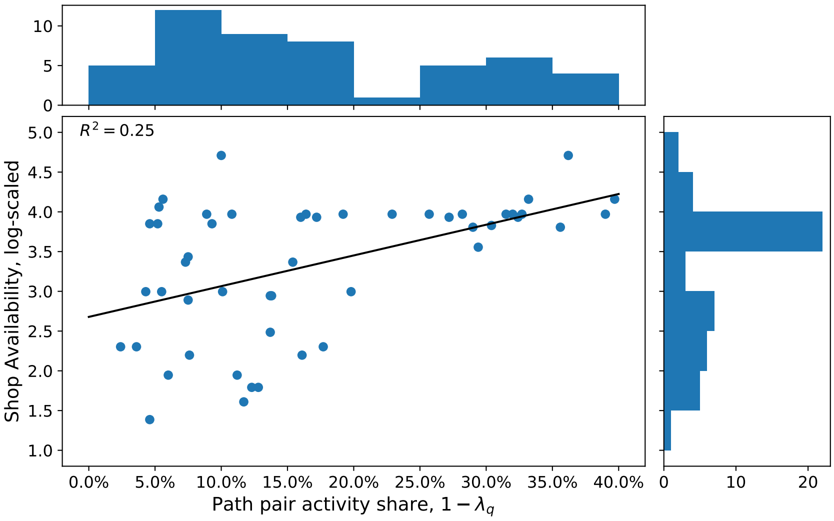


Figure 11: Results of shop availability and activity share relation. Only path-pairs with more than 2000 observations are included.

5.3. *Waiting times for different passenger groups*

Given the already identified feasible trains cf. Section 5.1 we extend this further to an actual train assignment by minimizing the exit time (i.e. $Tap\ Out_i - Train\ Arrival_{j,k}$). With the passenger trips assigned to trains it is possible to calculate the waiting time on the train platform. Since some trips has several feasible trains we have only focused on the trips with exactly one feasible train itinerary to limit the uncertainty of the true waiting time. Having these groups, the observed waiting and walking time can be plotted for each station as seen in Figure 12.

For Valby, the passengers predicted to the directly walking group have the lowest walking and waiting time compared to the activity group. The low walking and waiting time align with the assumption that the directly walking group describes the passengers who walk directly to minimize their overall transfer time. In the case of Korsør we see the same pattern for the walking time, with lowest mean walking time for the directly walking group and highest for activity group, but the reverse pattern for the waiting time. This indicates that the bus arrival and train departures are not synchronised, especially when taking into account that the median transfer time is 14.7 minutes for Korsør compared to Valby's 6.5 minutes. The lack of synchronisation between busses and trains make it difficult to minimize the overall transfer time for the directly walking passengers, which just results in a high waiting time. This shows that it is possible for the model to separate the activity of waiting in the station building and walking directly to train platform, thereby being able to identify inefficient connections.

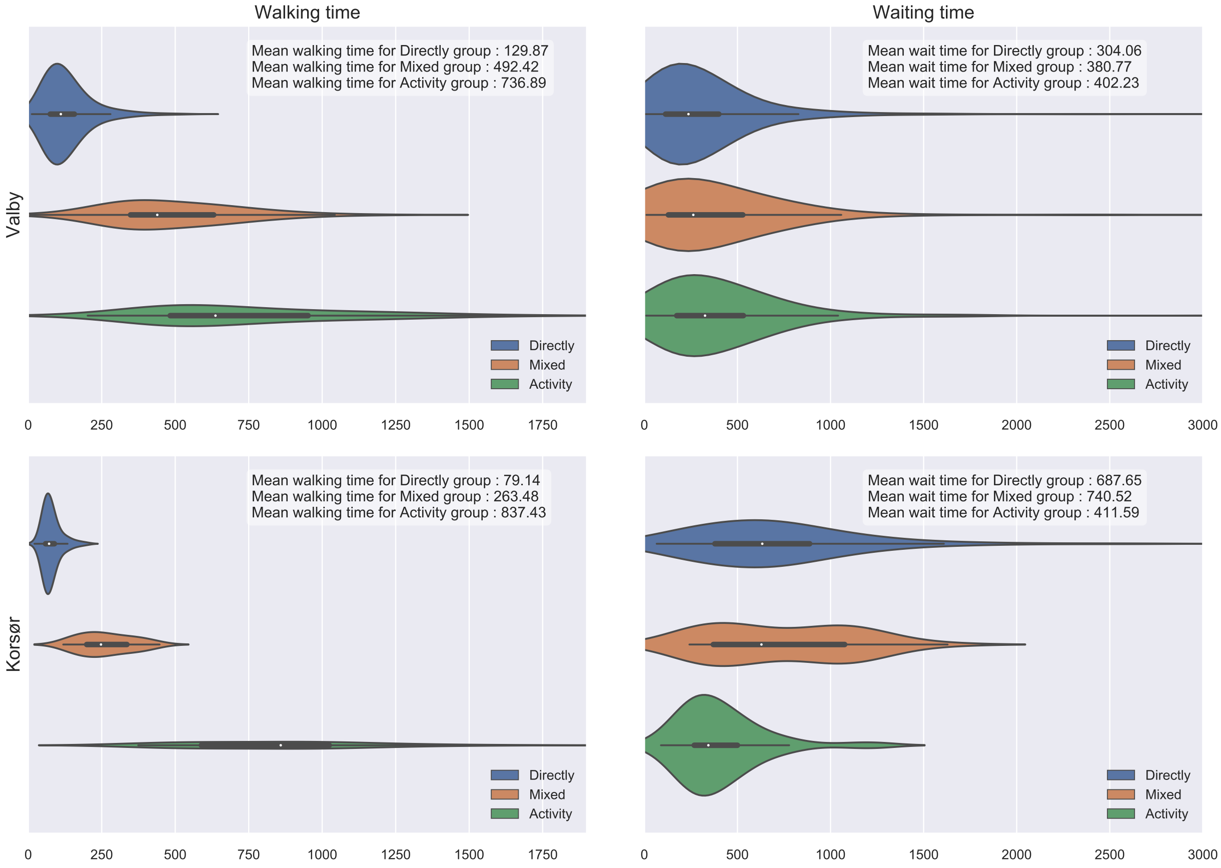


Figure 12: Observed waiting and walking time distribution for Valby and Korsør for each prediction group.

6. Conclusion

This study has presented a novel methodology for providing accurate walking time distributions at transfers from bus to train based on smart card data. The model requires AVL data from busses and smart card data where the passenger must tap-in at the train station, preferably at the platform to avoid uncertainty of possible time spent in a station building.

The proposed approach is able to reproduce the observed times between the passenger alights a bus taps in at the platform using a hierarchical Bayesian mixture model, where passengers are assumed to either walk directly to the platform or perform an activity during the transfer. The model is applied to a large-scale case study with 129 stations in the Eastern part of Denmark. Detailed investigations from two stations show that the model is able to estimate accurate walking time distributions for two types of stations: i) stations where passengers are spending extra time during the transfer due to poor synchronisation between busses and trains, and ii) stations where passengers or are doing shopping, buying coffee other short errands during the transfer.

The model can be easily applied at scale, and thus offer a more feasible methodology than manual surveys where passengers are followed through the transfer, when public transport agencies need to estimate the necessary walking time to perform transfers. The resulting distribution for walking time for the direct walking passengers can be compared to the scheduled walking time published by public transport agencies, and thereby identifying places where extra scheduled walking time is needed. In this way the agencies are able to plan more reliable connections between busses and trains by also including the variability of passengers' walking times.

Acknowledgements

The authors would like to thank Rail Net Denmark (BaneDanmark), Movia and DSB (Danish State Railways) for providing automatic vehicle location data for the project. Furthermore, the authors would like to thank Movia for providing data from the Danish smart card (Rejsekort) for use in the project.

Funding sources and declaration of interests

The authors of this work were partially funded by the Danish Innovation Foundation under grants '7038-00186B' and '4109-00005B'.

Declarations of interest: none

References

Anderson, M.K., 2013. Behavioural Models for Route Choice of Passengers in Multimodal Public Transport Networks. Ph.D. thesis. Technical University of Denmark.

Chen, Z., Zhao, X., Shi, R., 2016. Walking Speed Modeling on Transfer Passengers in Subway Passages, in: Proceedings of the International Conference on Civil, Transportation and Environment (ICCTE 2016), pp. 639–643. doi:10.2991/iccte-16.2016.107.

Daamen, W., Bovy, P.H., Hoogendoorn, S.P., 2006. Choices between stairs, escalators and ramps in stations, in: 10th International Conference on Computer System Design and Operation in the Railway and Other Transit Systems, COMPRAIL 2006, CR06, pp. 3–12. doi:10.2495/CR060011.

Daamen, W., Hoogendoorn, S.P., 2006. Free Speed Distributions for Pedestrian Traffic, in: Proceedings of the 85th Annual Meeting of the Transportation Research Board, pp. 13–25.

Dixit, M., Brands, T., van Oort, N., Cats, O., Hoogendoorn, S., 2019. Passenger Travel Time Reliability for Multimodal Public Transport Journeys. Transportation Research Record 2673, 149–160. doi:10.1177/0361198118825459.

Du, P., Liu, C., Liu, Z.L., 2009. Walking time modeling on transfer pedestrians in subway passages. Jiaotong Yunshu Xitong Gongcheng Yu Xinxi/ Journal of Transportation Systems Engineering and Information Technology 9, 103–109. doi:10.1016/s1570-6672(08)60075-6.

Faroqi, H., Mesbah, M., Kim, J., 2018. Applications of transit smart cards beyond a fare collection tool: a literature review. Advances in Transportation Studies: an international Journal 45, 107–122. doi:10.15713/ins.mmj.3.

Fruin, J.J., 1971. Pedestrian planning and design. Metropolitan Association of Urban Designers and Environmental Planners, New York.

Fujiyama, T., Cao, B., 2016. Lengths of time passengers spend at railway termini: An analysis using smart card data, in: 2016 IEEE International Conference on Intelligent Rail Transportation, ICIRT 2016, IEEE. pp. 139–144. doi:10.1109/ICIRT.2016.7588723.

GeoDanmark-data, 2020. Rights for usage of GeoDanmark-data. URL: https://www.geodanmark.dk/home/vejledninger/vilkaar-for-data-anvendelse/.

Google, 2020. GTFS Realtime Overview. URL: https://developers.google.com/transit/gtfs-realtime.

Guo, Z., Wilson, N.H., 2011. Assessing the cost of transfer inconvenience in public transport systems: A case study of the London Underground. Transportation Research Part A: Policy and Practice 45, 91–104. URL: http://dx.doi.org/10.1016/j.tra.2010.11.002, doi:10.1016/j.tra.2010.11.002.

Ingvardson, J.B., Nielsen, O.A., Raveau, S., Nielsen, B.F., 2018. Passenger arrival and waiting time distributions dependent on train service frequency and station characteristics: A smart card data analysis. Transportation Research Part C 90, 292–306. doi:10.1016/j.trc.2018.03.006.

Iseki, H., Taylor, B.D., 2009. Not All Transfers Are Created Equal: Towards a Framework Relating Transfer Connectivity to Travel Behaviour. Transport Reviews 29, 777–800. URL: http://www.tandfonline.com/doi/abs/10.1080/01441640902811304, doi:10.1080/01441640902811304.

Jang, W., 2010. Travel time and transfer analysis using transit smart card data. Transportation Research Record , 142–149doi:10.3141/2144-16.

Kasehyani, N.H., Abd Rahman, N., Abdul Sukor, N.S., Halim, H., Katman, H.Y., Abustan, M.S., 2019. Evaluation of pedestrian walking speed in rail transit terminal. International Journal of Integrated Engineering 11, 26–36. doi:2229838x.

Kouwenhoven, M., De Jong, G.C., Koster, P., Van Den Berg, V.A.C., Verhoef, E.T., Bates, J., Warffemius, P.M.J., 2014. New values of time and reliability in passenger transport in The Netherlands. Research in Transportation Economics 47, 37–49. doi:10.1016/j.retrec.2014.09.017.

Leurent, F., Xie, X., 2017. Exploiting smartcard data to estimate distributions of passengers' walking speed and distances along an urban rail transit line, in: Transportation Research Procedia, Elsevier B.V.. pp. 45–54. URL: http://dx.doi.org/10.1016/j.trpro.2017.03.006, doi:10.1016/j.trpro.2017.03.006.

Li, W., Yan, X., Li, X., Yang, J., 2020. Estimate Passengers' Walking and Waiting Time in Metro Station Using Smart Card Data (SCD). IEEE Access 8, 11074–11083. URL: https://ieeexplore.ieee.org/document/8954711/, doi:10.1109/ACCESS.2020.2965155.

Mosallanejad, M., Somenahalli, S., Vij, A., Mills, D., 2018. Distinguishing transfer from activity using public transport fare data, in: ATRF 2018 - Australasian Transport Research Forum 2018, Proceedings, pp. 1–5.

OpenStreetMap contributors, 2018. Denmark extract retrieved from https://download.geofabrik.de/europe/denmark.html. URL: https://www.openstreetmap.org.

Parbo, J., Nielsen, O.A., Prato, C.G., 2014. User perspectives in public transport timetable optimisation. Transportation Research Part C: Emerging Technologies 48, 269–284. doi:10.1016/j.trc.2014.09.005.

Pelletier, M.P., Trépanier, M., Morency, C., 2011. Smart card data use in public transit: A literature review. Transportation Research Part C 19, 557–568. doi:10.1016/j.trc.2010.12.003.

Raveau, S., Guo, Z., Muñoz, J.C., Wilson, N.H.M., 2014. A behavioural comparison of route choice on metro networks: Time, transfers, crowding, topology and socio-demographics. Transportation Research Part A 66, 185–195. doi:10.1016/j.tra.2014.05.010.

Rejseplanen (Danish Travel Planner), 2020. Rejseplanen.dk. URL: https://www.rejseplanen.dk.

Sarma, A., Kay, M., 2020. Prior Setting in Practice: Strategies and Rationales Used in Choosing Prior Distributions for Bayesian Analysis, in: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, ACM, New York, NY, USA. pp. 1–12. URL: https://dl.acm.org/doi/10.1145/3313831.3376377, doi:10.1145/3313831.3376377.

Schakenbos, R., La, L., Nijenstein, S., Geurs, K.T., 2016. Valuation of a transfer in a multimodal public transport trip. Transport Policy 46, 72–81.

Singh, R., Hörcher, D., Graham, D.J., Anderson, R.J., 2020. Decomposing journey times on urban metro systems via semiparametric mixed methods. Transportation Research Part C: Emerging Technologies 114, 140–163. doi:10.1016/j.trc.2020.01.022.

Sun, L., Jian Gang, J., Lee, D.H., Axhausen, K.W., 2015. Characterizing multimodal transfer time using smart card data The effect of time, passenger age, crowdedness and collective pressure, in: Transportation Research Board 94th Annual Meeting, pp. 1–15. URL: https://doi.org/10.3929/ethz-a-010025751.

Wahaballa, A.M., Kurauchi, F., Schmöcker, J.D., Iwamoto, T., 2018. Rail-to-Bus and Bus-to-Rail Transfer Time Distributions Estimation Based on Passive Data, in: Proceedings for the 14th Conference on Advanced Systems in Public Transport and Transit Data 2018, Brisbane, Australia. pp. 1–7.

Xiao, M., Chien, S., Hu, D., 2016. Optimizing coordinated transfer with probabilistic vehicle arrivals and passengers' walking time. Journal of Advanced Transportation 50, 2306–2322. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/atr.1460, doi:10.1002/atr.1460.

Xie, X., Leurent, F., 2017. Estimating distributions of walking speed, walking distance, and waiting time with automated fare collection data for rail transit. Transportation Research Record 2648, 134–141. doi:10.3141/2648-16.

Young, S.B., 1999. Evaluation of Pedestrian Walking. Transportation Research Record: Journal of the Transportation Research Board 1, 20–26.

Zhou, Y., Yao, L., Gong, Y., Chen, Y., 2016. Time prediction model of subway transfer. SpringerPlus 5. URL: http://www.springerplus.com/content/5/1/44, doi:10.1186/s40064-016-1686-7.

Zhu, W., Fan, W., Wei, J., Fan, W.D., 2020. Complete Estimation Approach for Characterizing Passenger Travel Time Distributions at Rail Transit Stations. Journal of Transportation Engineering, Part A: Systems 146. doi:10.1061/JTEPBS.0000375.

Zhu, Y., Koutsopoulos, H.N., Wilson, N.H., 2017. A probabilistic Passenger-to-Train Assignment Model based on automated data. Transportation Research Part B 104, 1–21. doi:10.1016/j.trb.2017.04.012.

Chapter 4

Identifying areas of interest

An aspect of the public transportation network not stored by the AFC system is the *areas of interest*, these being places that travellers want to travel to or from, but that are not served by public transportation. These places are of interest to public transit agencies since, if found, they can improve the attractiveness of the public transportation system and may even make it possible to increase the market share of public transportation users. This chapter investigates the challenges and the potential of combining smart card data with journey planner search data to discover the *areas of interest*.

Areas of interest

The *areas of interest* are not stored in the system since the locations of taps in and out for smart card trips are limited to the locations already in the public transportation network, and they do not show passengers' initial origins and final destinations. Fortunately, this limitation does not apply to search data from journey planners such as the Danish Rejseplanen. Rejseplanen users can search on addresses located outside the transportation network, thus revealing their final origin and destination locations. The drawback of the journey planner search data is that it does not show the actual demand from and to a location, since a traveller can search for the same trip several times, without ever actually undertaking the trip. In contrast to journey planner data, the smart card data observes demand to and from locations for smart card trips.

The underlying idea here is that the *areas of interest* can be identified if the relationship between journey planner searches and smart card trips can be isolated and learned.

Literature

To the best of the author's knowledge, limited attention has been given to combining smart card data with journey planner search data by the transportation research community. The exceptions are two Master's theses by Roosmalen (2019) and Wang (2020). Both theses investigate how journey planner search data, in combination with smart card data, can be used to forecast the travel demand of existing public transit routes.

In the following sections, an exploratory analysis of the data is performed to investigate the relation between the Danish smart card data (Rejsekort) and the Danish journey planner data (Rejseplanen).

4.1 Exploratory data analysis

We consider a data set provided by Rejsekort & Rejseplanen A/S that contains 139 million smart card trips with 340 million journey planner searches made during 2018. The smart card trips are spread over 22 thousand locations with 4.6 million different connections between the locations, and the journey planner searches are spread over 2.1 million locations with 34 million different connections. The unique locations for the smart card and journey searches are displayed in fig 4.1.

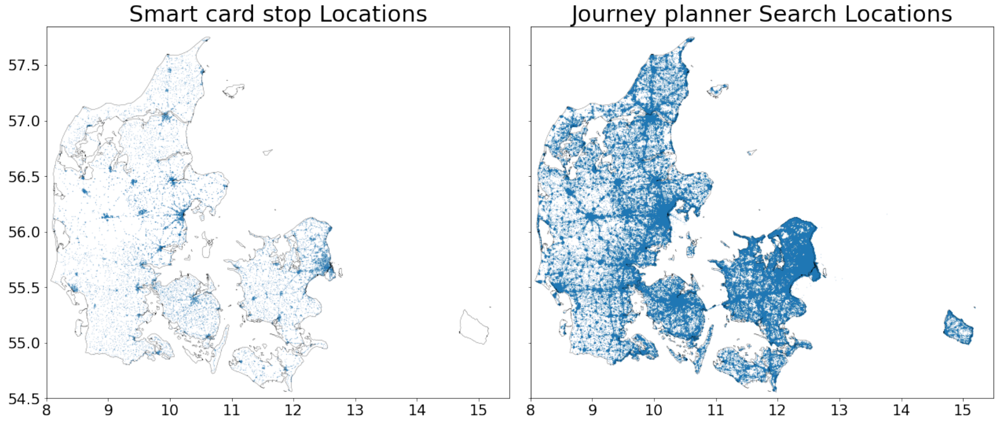


Figure 4.1: The Danish locations of smart card stops and journey planner searches.

4.1.1 Municipalities as areas

Due to the vast number of locations in the smart card trips and journey planner searches, the analysis can be simplified by defining the areas as municipalities. The subdivision by municipalities is valuable since the municipalities fund the public transportation routes inside and between them. This means that, if it possible to identify a municipality as an *area of interest*, it will be valuable not only for transit agencies but also for the municipalities themselves. Denmark is divided into 98 municipalities, of which five municipalities (Samsø, Ærø, Bornholm, Fanø, Læsø) do not use the Danish AFC system. These five are all islands, reachable only by boat or aeroplane. These five municipalities are excluded from the analysis.

Number of trips and searches to and from municipalities

Table 4.1 displays descriptive statistics for the smart card trips and journey planner searches *to* and *from* a municipality. Considering the averages over all municipalities, the table shows that the mean number of trips *from* a municipality is 1,493,670, and the median number of trips *from* a municipality is 568,276. This mean and median differ by a factor of nearly three, and the same can be seen for the mean and median for the number of trips *to* a municipality. For the number of searches, the pattern continues with the mean number of searches being nearly twice the median number of searches, both *from* and *to* a municipality. From these ratios, it can be deduced that distributions for the trips and searches *from* and *to* municipalities are all right-skewed. In addition, there are notably more journey planner searches than smart card trips. Taking the ratio between the number of searches and trips for each municipality, the number of searches is 3.5 times larger than searches *to* and *from* the same municipality. This notable difference between the number

Table 4.1: Descriptive statistics of the number of smart card trips and journey planner searches *to* and *from* municipalities.

	Smart card trips		Journey planner searches	
	From	To	From	To
Mean	1,493,670	1,493,598	3,631,083	3,629,874
Std	4,799,725	4,774,383	8,537,038	8,414,957
Min	30,710	31,096	191,656	170,288
25%	291,514	291,130	927,775	963,062
50%	568,276	570,464	1,914,256	1,782,062
75%	1,151,589	1,151,696	3,263,085	3,305,853
Max	45,187,722	44,928,034	77,279,142	74,499,930

of trips and searches for each municipality may have several reasons, such as:

- The traveller may search for the same trip several times with only one trip realised or with no realised trip at all.
- As mentioned in section 1.3, smart card trips constitute only a subset of all the trips in the public transportation network. Users performing a search for a given trip may not have used a smart card to pay for that trip, and hence, this particular trip will not show up in the smart card data despite appearing in the journey planner data.

4.1.2 The probability of searching and travelling to a municipality

As there were notably more searches than trips in each municipality, the trips and searches need to be rescaled to allow a more meaningful comparison between them. A simplistic approach to identify the *area of interest* is to compare the probability of taking a trip *to* and *from* a municipality with the probability of searching for a trip *to* and *from* a municipality.

Let M denote the set of municipalities, and let $t^{m\rightarrow}$ and $t^{\rightarrow k}$ denote trips travelling *from* municipality $m \in M$ and *to* municipality $k \in M$, respectively. With these, the probability of travelling *from* the municipality can be estimated by using the proportion of trips *from* the municipality out of the total number of trips travelling *from* all of the M municipalities, such that

$$P(T^{\text{From}} = m) = \frac{t^{m\rightarrow}}{\sum_{m'=1}^{M} t^{m'\rightarrow}}, \tag{4.1}$$

and for the probability of travelling *to* the municipality

$$P(T^{\text{To}} = k) = \frac{t^{\rightarrow k}}{\sum_{k'=1}^{M} t^{\rightarrow k'}}. \tag{4.2}$$

The same can be done for the probability of searching for trips *from* a municipality and probability of searching for a trips *to* a municipality by using the number of searches s, giving

$$P(S^{\text{From}} = m) = \frac{s^{m\rightarrow}}{\sum_{m'=1}^{M} s^{m'\rightarrow}} \tag{4.3}$$

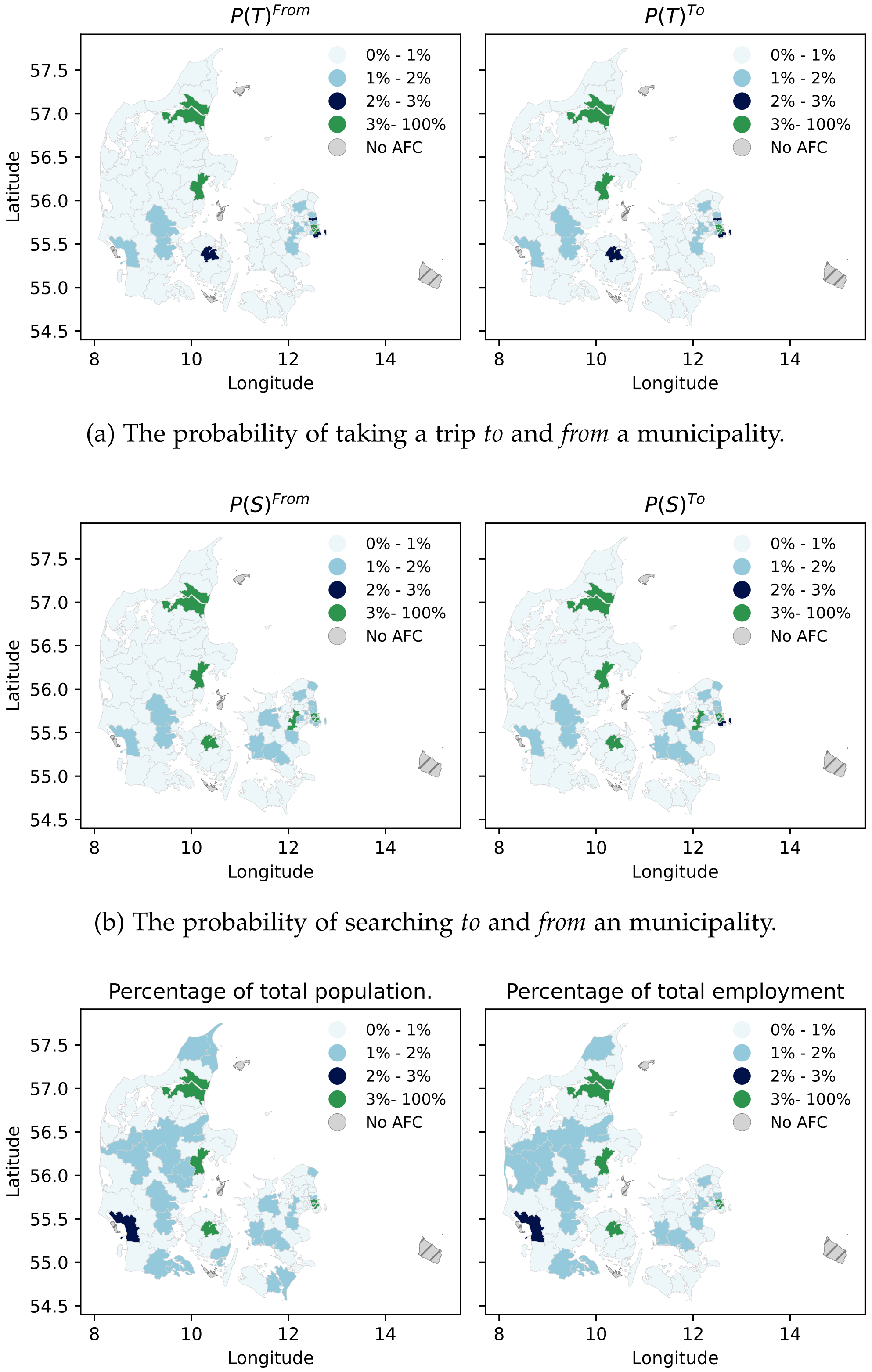
and

$$P(S^{\text{To}} = k) = \frac{s^{\rightarrow k}}{\sum_{k'=1}^{M} s^{\rightarrow k'}}. \tag{4.4}$$

In this case, there is no need for a Bayesian model since each municipality has a large number of samples, and there is therefore no need to regularise the quantities. In addition, both a maximum likelihood and Bayes model with weakly informed priors will give the same result. The four equations above are estimated with maximum likelihood due to that estimation being simpler to perform.

Description of probabilities

A geographical overview of the approximated probabilities is displayed in fig. 4.2, where fig. 4.2a shows the probability of taking a trip $\{P(T^d)\}_{d\in\{\text{From,To}\}}$ and 4.2b shows the probability of searching $\{P(S^d)\}_{d\in\{\text{From,To}\}}$. In both figure fig.4.2a and figure 4.2b, it is shown that the same municipalities are emphasised with higher probabilities. Many of these municipalities have larger share of the total population and employment levels, as shown in figure 4.2c. This becomes even clearer from the correlation matrix and the descriptive statistics of the probabilities in table 4.2. For the correlation in table 4.2a, the probabilities are highly correlated with each other, the municipalities' populations, and employment in the municipalities. In addition, when viewing the descriptive statistics in table 4.2b, it can be seen that the majority of municipalities have probabilities below 1%.



(a) The probability of taking a trip *to* and *from* a municipality.

(b) The probability of searching *to* and *from* an municipality.

(c) The percentage that each municipality contributes to the total in relation to the population and employment demographics (Danmarks Statisk, 2021).

Figure 4.2: The probability of taking a trip and the probability of searching for trips from and to a municipality, with demographics.

Table 4.2: Correlation matrix and descriptive statistics for the probability of taking a trip and the probability of searching for a trip from and to a municipality, with the population and employment demographics.

(a) Correlation matrix

	Smart card trips		Journey planner searches		Of total	
	$P(T^{\text{From}})$	$P(T^{\text{To}})$	$P(S^{\text{From}})$	$P(S^{\text{To}})$	Employment	Population
$P(T^{\text{From}})$	1.000	1.000	0.973	0.962	0.935	0.907
$P(T^{\text{To}})$	-	1.000	0.973	0.963	0.935	0.908
$P(S^{\text{From}})$	-	-	1.000	0.997	0.973	0.954
$P(S^{\text{To}})$	-	-	-	1.000	0.972	0.955
Employment	-	-	-	-	1.000	0.989
Population	-	-	-	-	-	1.000

(b) Descriptive statistics

	Smart card trips		Journey planner searches		Of total	
	$P(T^{\text{From}})$	$P(T^{\text{To}})$	$P(S^{\text{From}})$	$P(S^{\text{To}})$	Population	Employment
Mean	1.07%	1.07%	1.07%	1.07%	1.07%	1.07%
Std	3.45%	3.44%	2.52%	2.49%	1.26%	1.62%
Min	0.02%	0.02%	0.06%	0.05%	0.22%	0.11%
25%	0.21%	0.21%	0.27%	0.28%	0.57%	0.46%
50%	0.41%	0.41%	0.57%	0.53%	0.78%	0.74%
75%	0.83%	0.83%	0.96%	0.98%	1.06%	1.06%
Max	32.52%	32.33%	22.85%	22.03%	10.61%	14.01%

The seven largest contributors to probability mass

The low probabilities associated with the majority of municipalities are due to seven municipalities accounting for a large share of the probability mass. In the case of the trips, these seven municipalities account for 56% of the probability mass for *from* municipalities, and 55% *to* municipalities. In the case of the searches, they account for 47% *from* municipalities and 46% *to* municipalities. Four of the seven municipalities are Denmark's most populated. The largest municipality, København, where the capital is located, has the maximum in all four probabilities. Of the three left, the seventh largest municipality, Frederiksberg, is unique since it lies inside the municipality of København, and is the only municipality with more people travelling to it, than searching for travel to it.

The last two of the seven municipalities differ, in terms of the various probabilities. For the probabilities $\{P(T^d)\}_{d\in\{\text{From,To}\}}$ it is the same municipalities Lyngby-Taarbæk and Tårnby, whereas for $P(S^{\text{Form}})$ has the municipalities Roskilde and Høje-Taastrup and $P(S^{\text{To}})$ have the municipalities Roskilde and Tårnby. Tårnby's high probability may be explained by its having the largest airport in Denmark. The high probabilities of Roskilde and Lyngby-Taarbæk could be explained by both containing larger universities. The last municipality, Høje-Taastrup, is the thirty-eighth largest municipality and contains the last stop for direct lines from the capital to the major cities on Fyn and Jutland. The probabilities of all municipalities with their population and employment demographics can be found in appendix B.1 table B.1.

4.1.3 The probability ratio

Having the probabilities $\{P(T^d)\}_{d\in\{\text{From,To}\}}$ and $\{P(S^d)\}_{d\in\{\text{From,To}\}}$, the *area of interest* can be quantified by constructing ratios i.e.

$$\text{Ratio}_m^{\text{From}} = \frac{P(\text{S}^{\text{From}} = m)}{P(\text{T}^{\text{From}} = m)} \tag{4.5}$$

and

$$\text{Ratio}_k^{\text{To}} = \frac{P(\text{S}^{\text{ To}} = k)}{P(\text{T}^{\text{To}} = k)}. \tag{4.6}$$

By using a ratio to quantify the *areas of interest*, the hypothesis is that higher ratios will indicate places that people want to travel to or from, but that are not served by public transportation.

The $\{\text{Ratio}_m^d\}_{d\in\{\text{From,To}\}}$ of each municipality is displayed in figure 4.3, showing it is the same municipalities that have the highest ratios *to* and *from* them. In addition, these municipalities are clustered together on the map. The top ten highest $\{\text{Ratio}_m^d\}_{d\in\{\text{From,To}\}}$ belong to municipalities with a probability $\{P(S^d)\}_{d\in\{\text{From,To}\}}$ 2.5 times the size of $\{P(T^d)\}_{d\in\{\text{From,To}\}}$, and are presented in table 4.3. These ten municipalities are potential *areas of interest* since they are the municipalities with the highest discrepancies between $\{P(T^d)\}_{d\in\{\text{From,To}\}}$ and

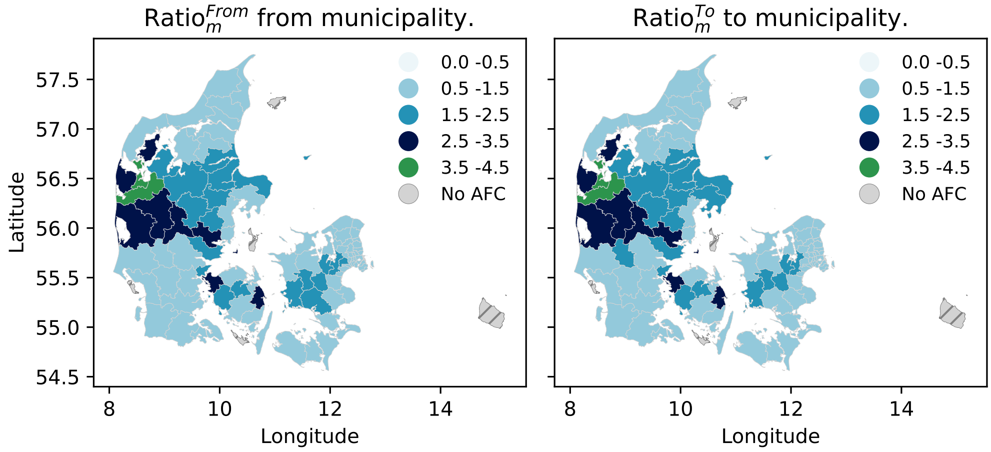


Figure 4.3: The ratio between the probability of taking a trip and the probability of searching from and to an municipality.

Table 4.3: The $\{\text{Ratio}_m^d\}_{d\in\{\text{From,To}\}}$ and probability of the taking a trip and searching to and from a municipality with the corresponding the number of trips and searches for the top ten municipalities with highest $\{\text{Ratio}_m^d\}_{d\in\{\text{From,To}\}}$.

	Trips		Searches		P(T)		P(S)		Ratio	
Municipality	From	To	From	To	From	To	From	To	From	To
Holstebro	184,897	185,312	1,676,721	1,782,062	0.13%	0.13%	0.50%	0.53%	3.73	3.95
Struer	91,525	91,149	820,042	816,629	0.07%	0.07%	0.24%	0.24%	3.68	3.68
Middelfart	244,094	242,767	2,038,750	1,939,827	0.18%	0.17%	0.60%	0.57%	3.43	3.28
Herning	409,942	414,021	2,979,692	3,305,853	0.29%	0.30%	0.88%	0.98%	2.99	3.28
Ikast-Brande	104,763	105,004	879,977	827,466	0.08%	0.08%	0.26%	0.24%	3.45	3.24
Horsens	409,097	414,265	3,263,085	3,070,425	0.29%	0.30%	0.96%	0.91%	3.28	3.05
Ringkøbing-Skjern	157,222	156,605	1,157,771	1,118,703	0.11%	0.11%	0.34%	0.33%	3.03	2.94
Lemvig	32,828	33,493	217,486	222,454	0.02%	0.02%	0.06%	0.07%	2.72	2.73
Nyborg	382,638	378,575	2,738,725	2,457,430	0.28%	0.27%	0.81%	0.73%	2.94	2.67
Morsø	30,710	31,096	191,656	190,013	0.02%	0.02%	0.06%	0.06%	2.56	2.51

$\{P(S^d)\}_{d\in\{\text{From,To}\}}$. However, it is difficult to assess whether these municipalities are locations that people want to travel *to* and *from*, and whether they are areas where connections to the public transportation network are sub-optimal. Another reason for discrepancies could be that people are more likely to search *to* and *from* these municipalities because people less frequently take trips *to* and *from* these municipalities than other municipalities.

Conditioning on the municipality.

To improve the understanding of the underlying patterns in the trips and searches *to* and *from* a specific municipality, the conditional probability is approximated by a simple Bayesian model. There are two main conditions: the first is the condition of starting from the municipality m *to* the M municipalities, denoted $m \rightarrow *$, and the second condition is going *to* the municipality k *from* the M municipalities, denoted $* \rightarrow k$. E.g. the number trips condition on starting *from* the municipality m and going to the M municipalities will be: $\boldsymbol{t}^{m \rightarrow *} = [t^{m\rightarrow 1}, t^{m \rightarrow 2}, \ldots, t^{m \rightarrow M}]$.

For simplicity, it is assumed that the probability of taking a trip $\{\boldsymbol{\pi}^d\}_{d \in \{m \rightarrow *, * \rightarrow k\}}$ and the probability of searching $\{\boldsymbol{\lambda}^d\}_{d \in \{m \rightarrow *, * \rightarrow k\}}$ are independent of each other, and that the municipalities are independent of each other. This assumption is obviously crude, however, it suffices as a simple starting point and makes it possible to infer the probabilities. Further, it is assumed that the probabilities of taking a trip $\{\boldsymbol{\pi}^d\}_{d \in \{m \rightarrow *, * \rightarrow k\}}$ and searching $\{\boldsymbol{\lambda}^d\}_{d \in \{m \rightarrow *, * \rightarrow k\}}$ both follow a multinomial distribution with a conjugated Dirichlet prior with the hyperparameters $\{\boldsymbol{\alpha}^d\}_{d \in \{m \rightarrow *, * \rightarrow k\}}$ for $\{\boldsymbol{\pi}^d)\}_{d \in \{m \rightarrow *, * \rightarrow k\}}$ and $\{\boldsymbol{\beta}^d\}_{d \in \{m \rightarrow *, * \rightarrow k\}}$ for $\{\boldsymbol{\lambda}^d\}_{d \in \{m \rightarrow *, * \rightarrow k\}}$.

With this, the conditional probabilities for the trips are

$$P(\boldsymbol{\pi}^{* \rightarrow k} | \boldsymbol{t}^{* \rightarrow k}, \boldsymbol{\alpha}^{* \rightarrow k}) \propto \mathrm{Multi}(\boldsymbol{t}^{* \rightarrow k} | \boldsymbol{\pi}^{* \rightarrow k}) \mathrm{Dir}(\boldsymbol{\pi}^{* \rightarrow k} | \boldsymbol{\alpha}^{* \rightarrow k}) \tag{4.7}$$

and

$$P(\boldsymbol{\pi}^{m \rightarrow *} | \boldsymbol{t}^{m \rightarrow *}, \boldsymbol{\alpha}^{m \rightarrow *}) \propto \mathrm{Multi}(\boldsymbol{t}^{m \rightarrow *} | \boldsymbol{\pi}^{m \rightarrow *}) \mathrm{Dir}(\boldsymbol{\pi}^{m \rightarrow *} | \boldsymbol{\alpha}^{m \rightarrow *}), \tag{4.8}$$

and the conditional probabilities of searching are

$$P(\boldsymbol{\lambda}^{* \rightarrow k} | \boldsymbol{s}^{* \rightarrow k}, \boldsymbol{\beta}^{* \rightarrow k}) \propto \mathrm{Multi}(\boldsymbol{s}^{* \rightarrow k} | \boldsymbol{\lambda}^{* \rightarrow k}) \mathrm{Dir}(\boldsymbol{\lambda}^{* \rightarrow k} | \boldsymbol{\beta}^{* \rightarrow k}) \tag{4.9}$$

and

$$P(\boldsymbol{\lambda}^{m \rightarrow *} | \boldsymbol{s}^{m \rightarrow *}, \boldsymbol{\beta}^{m \rightarrow *}) \propto \mathrm{Multi}(\boldsymbol{s}^{m \rightarrow *} | \boldsymbol{\lambda}^{m \rightarrow *}) \mathrm{Dir}(\boldsymbol{\lambda}^{m \rightarrow *} | \boldsymbol{\beta}^{m \rightarrow *}). \tag{4.10}$$

Due to the conditioning on the municipality, the data is partitioned into smaller subsets, and there will be occasions where only a few travellers make a trip to and from different municipalities. Therefore, a weakly informed prior of $\boldsymbol{\alpha} = [100]_{k=1}^{M}$ and $\boldsymbol{\beta} = [100]_{m=1}^{M}$ is used to avoid overfitting these cases. The hyperparameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ can be seen as pseudo-trips and searching, i.e. all municipalities are seen as having searches and travellers to and from them 100 times before seeing any data.

Once the probabilities $\{\boldsymbol{\pi}^{d}\}_{d\in\{m\to *,*\to k\}}$ and $\{\boldsymbol{\lambda}^{d}\}_{d\in\{m\to *,*\to k\}}$ have been determined, the discrepancies between municipalities can be found by taking the conditional ratios between these quantities, i.e.

$$\text{Ratio}_{k}^{m\to *} = \frac{\widetilde{\lambda}_{k}^{m\to *}}{\widetilde{\pi}_{k}^{m\to *}} \tag{4.11}$$

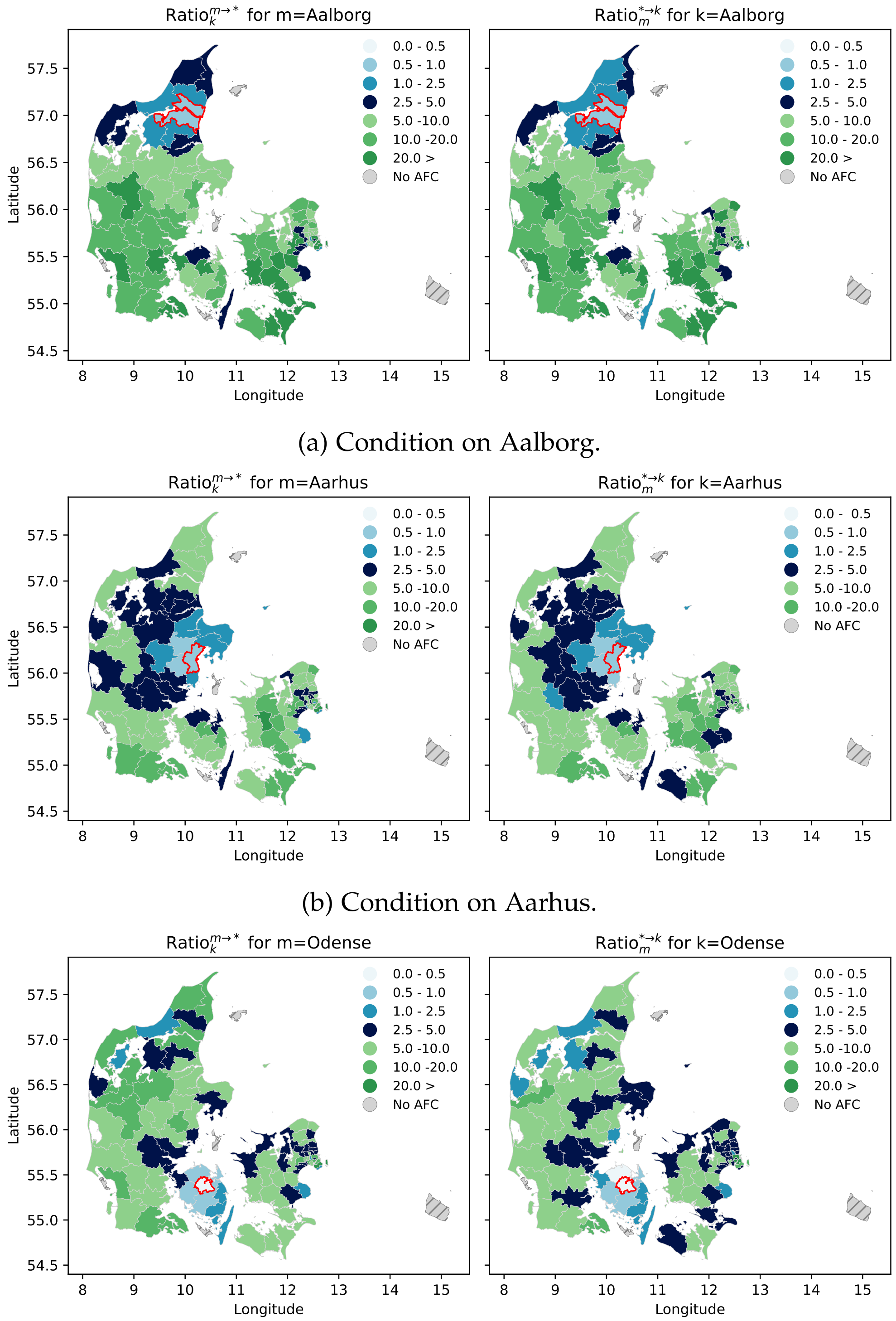
and

$$\text{Ratio}_{m}^{*\to k} = \frac{\widetilde{\lambda}_{m}^{*\to k}}{\widetilde{\pi}_{m}^{*\to k}}. \tag{4.12}$$

where $\tilde{\lambda}^{m\to *}$ is an estimator for the probability. In the following, the maximum a-posteriori (MAP) estimator is used, and similarly for the other probabilities.

The conditional ratio and potential *areas of interest*.

In fig. 4.4 and 4.5 the conditional ratios for the four largest municipalities and the municipality of Tårnby, which contains Denmark's largest airport, are displayed on a map of Denmark. These figures reveals that the distance between the municipalities affects the conditional ratio, an effect supported by fig. 4.6 showing the relationship between the Euclidean distance between two municipalities and, respectively, the number of trips and the number of searches. In both cases, the number of trips and searches decrease as a function of the Euclidean distance, and the number of trips falls notably faster than the number of searches. This pattern also appears when reviewing the municipalities identified as potential *areas of interest* in section 4.1.3 above, which are listed in detail in appendix B. In addition, these municipalities all have København and Tårnby in the top five highest conditional ratios, except Lemvig and

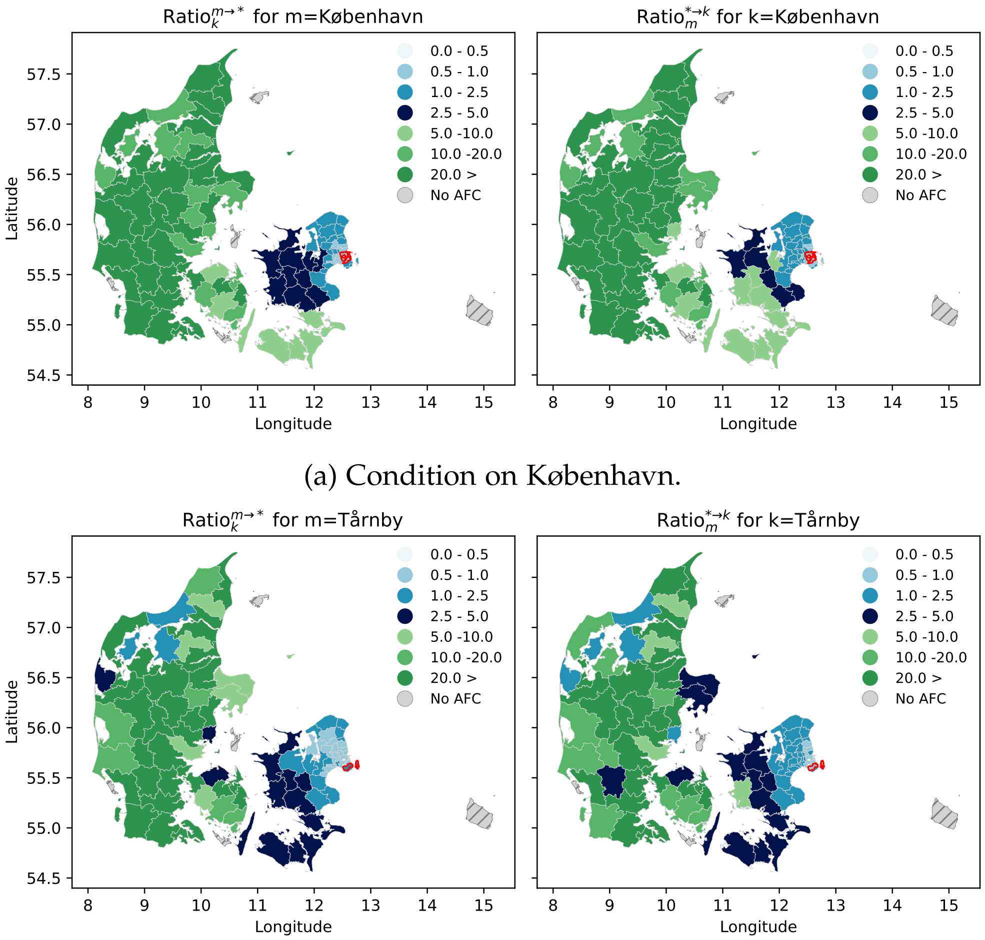

Ratio$_k^{m\to *}$ for m=Aalborg
Ratio$_m^{*\to k}$ for k=Aalborg
Ratio$_k^{m\to *}$ for m=Aarhus
Ratio$_m^{*\to k}$ for k=Aarhus
Ratio$_k^{m\to *}$ for m=Odense
Ratio$_m^{*\to k}$ for k=Odense
0.0 - 0.5
0.5 - 1.0
1.0 - 2.5
2.5 - 5.0
5.0 -10.0
10.0 -20.0
20.0 >
No AFC
Latitude
57.5
57.0
56.5
56.0
55.5
55.0
54.5
Longitude
8
9
10
11
12
13
14
15

(a) Condition on Aalborg.

(b) Condition on Aarhus.

(c) Condition on Odense.

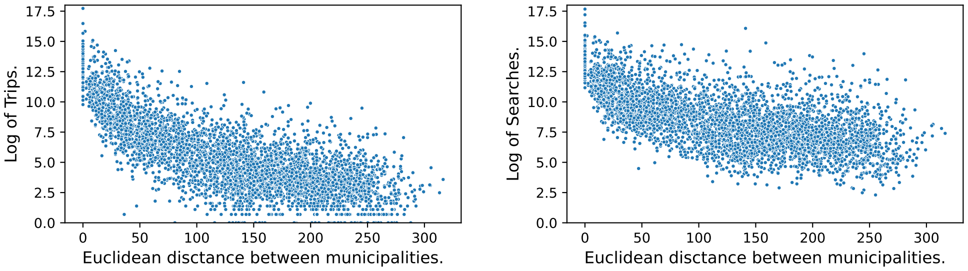
Figure 4.4: Ratio between the conditional probability of travelling and searching in relation to Aalborg, Aarhus and Odense. The municipality, which is condition on is outlined in red.


Ratio$_k^{m\to *}$ for m=København
Ratio$_m^{*\to k}$ for k=København
Ratio$_k^{m\to *}$ for m=Tårnby
Ratio$_m^{*\to k}$ for k=Tårnby
0.0 - 0.5
0.5 - 1.0
1.0 - 2.5
2.5 - 5.0
5.0 -10.0
10.0 -20.0
20.0 >
No AFC
Latitude
57.5
57.0
56.5
56.0
55.5
55.0
54.5
Longitude
8
9
10
11
12
13
14
15

(a) Condition on København.

(b) Condition on Tårnby.

Figure 4.5: Ratio between the conditional probability of travelling and searching in relation of København and Tårnby. The municipality, which is condition on is outlined in red.



(a) Number of trips to the direct distance between the municipalities.

(b) Number of searches to the direct distance between the municipalities.

Figure 4.6: The number of trips and searches between municipalities in relation to the direct distance between the municipalities.

Table 4.4: The conditional probabilities on going from the municipality and the ratio $\text{Ratio}_m^{*->k}$ with their corresponding populations, numbers of trips and searches for Lemvig and Morsø.

m	k	Population	Trips	Searches	$\tilde{\pi}_k^{m->*}$	$\tilde{\lambda}_k^{m->*}$	$\text{Ratio}_k^{m->*}$
Morsø	Morsø	20,514	8,788	67,585	22.21%	33.69%	1.52
Lemvig	Lemvig	20,133	17,191	78,825	41.04%	34.82%	0.85

Table 4.5: The conditional probabilities on going to the municipality and the ratio $\text{Ratio}_m^{*->k}$ with their corresponding population, number of trips and searches for Lemvig and Morsø.

m	k	Population	Trips	Searches	$\tilde{\pi}_m^{*->k}$	$\tilde{\lambda}_m^{*->k}$	$\text{Ratio}_m^{*->m}$
Morsø	Morsø	20,514	8,788	67,585	22.00%	33.96%	1.54
Lemvig	Lemvig	20,133	17,191	78,825	40.41%	34.07%	0.84

Morsø, which do not have Tårnby at the top. These two municipalities are noteworthy since they are both small municipalities with a population of around 20 thousand people, as can be seen in table 4.4 and 4.5. The table shows the conditional probabilities and ratio *to* and *from* the same municipality with their corresponding populations, and their numbers of trips and searches for Lemvig and Morsø. The thing to note is that the conditional probabilities of searches $\{\boldsymbol{\lambda}^d\}_{d\in\{m\to*,*\to k\}}$ are the same for both municipalities—between 33–34%, while the conditional probabilities of taking a trip $\{\boldsymbol{\pi}^d\}_{d\in\{m\to*,*\to k\}}$ are between 40–41% for Lemvig and 22% for Morsø. This interesting, since the municipalities are of the same size, and the conditional probabilities are the same $\{\boldsymbol{\lambda}^d\}_{d\in\{m\to*,*\to k\}}$, yet the probabilities of $\{\boldsymbol{\pi}^d\}_{d\in\{m\to*,*\to k\}}$ are notably different, and none of the ratios indicates notable discrepancies.

4.2 Discussion

The preliminary results above reveal that there are some municipalities with notably higher ratios, which could imply that they are *areas of interest*. The interpretation, however, is debatable. First of all, the lack of validation methods or ground truth for the *areas of interest* means that the results cannot be verified. Without verification, it is not straightforward to assess whether smart card data and journey planner searches are suitable for identifying *areas of interest*. Since some journeys may not be searched for, for instance by people who know the public transportation routes well, those travel desires will not feature in the search data, limiting search data's suitability for identifying *areas of interest*. There may be other factors, however, that explaining the higher ratios.

One such factor could be a behavioural factor tied to the types of trips that are searched for, and the types not searched for. It would be reasonable to assume that longer trips would be more likely to be searched for than shorter trips, which could explain the correlation between distance travel and higher ratios. The reason behind these trips being more likely could be:

- It would be expected that longer trips are more schedule-dependent, in the sense that missed departures would have a larger effect on the overall journey time of the trip. Therefore, the traveller would be incentivised to search the trip several times to ensure their departure. Since longer trips are less travelled than other trips, the traveller is less familiar with the journey. It would be expected that less regular trips, such as longer trips, will be searched for more than regular trips.
- It can be assumed that longer trips would tend to require more transfers and therefore be more complex for the traveller. The higher complexity of the trip would be expected to result in a higher number of searches than for simpler trips.

Given these reasonable additional factors, it could be argued that incorporating these factors into the models would create more realistic models for inferring *areas of interest*. Even with these additional factors it would still be difficult to determine a threshold for ratios to indicate an *area of interest* without some verification method.

Another approach would be to construct a model for the probability of taking a trip conditional on the trip being searched for. With this probability, it would be possible to identify trips with a high probability of being searched for and a low probability of being actually travelled. However, this requires a data set where individual trips can be associated with corresponding searches. In addition, the segmentation into municipalities may not be fine-grained enough to identify the *areas of interest*. The *areas of interest* could be more local, like local sightseeing sites or small companies where people work or have appointments. These nuances may be captured if the segmentation is smaller, such as city, segmented on land use zoning, or segmented based on the distance from the search address to the nearest entrance of the transportation network.

Even though the results are inconclusive, and it may be difficult to construct methods for verifying the results, further research is encouraged due to the possible benefits for public transportation. Firstly, a successful method would make it possible for public transit agencies to identify new areas where public transportation is in demand, thus making it possible for the agencies to increase ridership and revenue. Secondly, if the *areas of interest* are then connected to a transportation network, it would increase the attractiveness of public transportation by making it possible for people to travel by public transit to areas where they want to go.

Summary of the main observations in the chapter

In this exploratory study, the use of the smart card and online journey planner has been explored, resulting in the following observation with some more expected than others.

- There are notable differences in the number of trips and the number of searches, which can be explained by various reasons such as travellers searching for the same trip several times.
- There are strong correlations between probability of searches, the probability of travelling and the populations of municipalities.
- Highly populated municipalities account for a large fraction of the probability mass for both searches and trips.
- The probability of searching and the probability of travelling may be affected by large institutions such as universities or airports.
- The distances between municipalities correlate with the numbers of searches and trips made, and the number of trips falls notably faster then number of searches as a function of the Euclidean distance.
- The constructed ratios using the marginal and conditional probabilities both show a subset of municipalities with prominently higher ratios. However, it is difficult to access how much of that prominence indicates an *area of interest* and how much is due to other factors.

4.3 Conclusion

In this chapter, the possibility of combining smart card data with journey planner search data to identify *areas of interest* have been investigated. The study examined Danish municipality areas through smart card trips recorded in the Danish AFC system *Rejsekort* and searches in the Danish journey planner *Rejseplanen*. To quantify the *areas of interest*, the study explores the use of ratios between the smart card data and the journey planner searches, where the hypothesis is that higher ratios indicate places where people want to travel to or from but that are not served by the public transportation network.

Four ratios are inferred in the study, with two using the marginal probability and two using the conditional probability. For the ratios calculated from marginal probability, the first is the ratio between the probability of travelling *from* a municipality and the probability of searching *from* a municipality. The second ratio is between the probability of travelling *to* a municipality and the probability of searching *to* a municipality. The two other ratios are inferred by conditioning the probabilities respectively on starting *from* a municipality and going *to* a municipality.

The inferred ratios reveal that there are municipalities with higher ratios, which could indicate potential *areas of interest*. Nevertheless, other factors may contribute to the municipalities having higher ratios. Given this and the lack of ground truth for validating the proposed method, the results are thus far inconclusive. Despite this, the exploratory study shows a connection between the smart card data and journey planner search data, which can be a basis for further research.

4.3.1 Future Research

The above findings on the combination and use of the smart card data and online journey planner searches, through their limitations, open great opportunities for further research. Below are several relevant questions for further research directions.

- How can methods be developed or data sets obtained to validate methods for inferring the *area of interest*? This is important, since researchers and public transit agencies needs a way to assess the methods usefulness, before the method can included into their applications.

- How can the user performing the search be identified such that the same search, by the same user, does not appear several times? In addition, connecting the individual searches with the observed trips would give a more accurate picture of the relationship between the smart card trips and journey planner searches.

- Would more realistic models, accounting for the behavioural factors regarding when and what types of journeys people search for, lead to more insightful ratios? Incorporating these factors into the model and adjusting for their effect on the ratio would presumably give a better foundation for evaluating the use of ratios.

- Will the probability of taking a trip, conditional on the trip being searched for, yield better results? This would presumably be a better way of identifying the *area of interest*, since it would connect a trip with its associated searches. However, using conditional probability requires knowledge of the connections between the trip and its searches at the individual level.

- Could clustering, based on the distance between stop locations of trips and locations of the searches be used, instead of segmenting on municipalities? By connecting the trips and searches in this way, it may be possible to gain a better understanding of how the distance between locations of a search and the stop locations affects the number of trips in different areas.

Chapter 5

Conclusion

This PhD thesis has explored missing information in AFC data originating from the design challenges, and investigated how new Bayesian methods can infer that missing information.

Missing information can originate from different categories of challenge in the AFC system. The focus of this thesis has been on the missing information originating from the design challenge. Compared to the hardware, user and input challenges, where only a subset of the data is affected and can usually be handled by removing the affected data, the missing information rooted in the design challenge is systematic and affects the entire data set. Since the design challenge cannot be handled by removing the affected data, methods for inferring the missing information are essential for creating an accurate and sufficient description of the public transportation network.

Paper A explored the effect of missing information and showcased how it can induce errors in downstream analysis. The effects of missing recorded timetables are examined by using the scheduled timetable as a substitute for the recorded timetable. The analysis shows that the use of the scheduled timetables instead of the recorded timetables can induce notable errors in the passenger-to-train assignments and the tap-in and tap-out distribution of travellers. This showcases the importance of developing a method for inferring the missing information, since the public transit agencies or research community may not have access to the relevant data sources.

Paper A and B show how hierarchical Bayesian mixture models can infer the missing information of interest by utilising the domain knowledge of the problem, and chapter 4 explores how the missing *area of interest* may be infer using smart card data and online journey search data.

In paper A, the knowledge of how the tap-outs are distributed in relation to the train arrival times (Hong et al. 2016; Min et al. 2016; Tan et al. 2021) and how the train delays are distributed (Cerreto et al., 2018) are built into the model. With this knowledge embedded into the model, it is possible to infer the missing arrival times of trains with an average error of 30 to 42 seconds, depending on the station. The paper can impact how scheduled timetables are used in downstream analysis, which may be used to make decisions regarding the transportation network, decisions that may have important societal impacts. When public transit agencies evaluate the capacities of their trains, they need load profiles for the different trains to access whether they are over or under capacity. Using the scheduled timetable would, in some cases, lead to the wrong trains being classified as over- and under-capacity. Using these mis-classifications to regulate capacity may affect the comfort of travellers using these trains, leading some travellers to choose other forms of transportation.

In paper B, the knowledge that some travellers may perform activities (like shopping or buying coffee etc.) during their transfer, thereby having longer transfer times than travellers walking directly, are embedded into the model. This creates a model that can infer the missing direct walking time distributions during a transfer for 129 stations with 1,009 paths at the bus-to-platform level, making it possible for transit agencies to identify sub-optimal connections. Since the ground truth is not available, two validation methods were developed, which support the results of the proposed model. The models and the results could impact public transportation in Denmark. Currently, the Euclidean distance is used to access the scheduled walking time. This is done by each Danish public transit agency and delivered to Rejsekort & Rejseplanen A/S, which manage the national journey planner Rejseplanen and provides scheduled timetable information to other journey planners such as Google Maps and Apple Maps. The method could be implemented by each transit agency, given local improvements to the connections between transportation modes. However, if the company Rejsekort & Rejseplanen A/S, which also manages the Danish AFC system, implemented the proposed model, it would have a national impact. This could improve the provision of travel information to thousands of travellers, thus generating an impactful social benefit.

The study in chapter 4 explored the use of combining smart card data with online journey search data to infer underserved *areas of interest*. This study examined travel in 98 Danish municipalities through records of 138 million smart card trips and 340 million online journey searches from 2018. The exploration lead to the hypotheses of identifying the *areas of interest* using ratios between the probability of searching and the probability of taking a trip to and from a municipality. The hypothesis was that a higher ratio would indicate the *area of interest*. The results were inconclusive. With limited research on combining and using the smart card with the online journey planner data, the proposed methods had limited knowledge to incorporate into models.

Nevertheless, the gains on identifying *areas of interest* are potentially noteworthy. If these areas can be identified, transit agencies would be able to expand their operation to places for which there is an unmet public transport demand, thus improving the attractiveness of public transportation. Such demand-targetted expansion could lead to an uptake in ridership and revenue for the agencies, in addition to making it possible for people where they want.

In summary, the thesis has explored the missing information originating in the design challenge, and shown that, with domain knowledge, a

Bayesian framework can be used to infer missing information in the AFC system. The thesis has also contributed new methods for inferring missing information originating in the design challenge.

5.1 Reflections and future work

5.1.1 Paper A

Paper A showed measurable improvements for inferring the missing arrival times of trains. Nevertheless, the model relies on large assumptions, and the performance of the model could potentially be improved by relaxing these. The delays at each station are assumed to be independent of each other. This assumption is most likely invalid, since a train delay can persist over several successive stops. A more realistic approach would be to add a hierarchical layer which models the delays between stations. This additional layer would presumably improve the model by allowing it to use knowledge from multiple stations to infer train delays elsewhere. A further improvement would be to incorporate more behavioural knowledge into the model, such as whether people are travelling during peak or off-peak periods. A reasonable assumption is that there are more travellers during peak times than off-peak, meaning that it is more likely to observe train arrival times during peak periods. The Dirichlet prior could have incorporated this knowledge by making the trains in the peak periods more likely than those in the off-peak period. The knowledge could also have been incorporated into Dirichlet prior using additional data sources, such as the Danish National Travel Survey in the case of Denmark, to construct an empirical prior, or through direct incorporation into the model. A limitation of the model is the assumption that trains do not overtake each other, which limits applicability to lines without capacity for overtaking. Given that many public transit agencies do operate lines with overtaking, the model will not be accurate. These lines would be expected to be main lines, where the public transit agencies have an operational AVL system with high-quality data. However, on smaller and more local lines, the assumption of no-overtaking may hold, and as these lines are less likely to have enjoyed the same AVL systems investment, their associated AVL data is of lower quality. As a further research direction, it would be relevant investigating whether integrating these more realistic assumptions would improve the results and the applicability to lines where trains can overtake each other.

5.1.2 Paper B

The proposed model in paper B assumes time-invariant walking behaviour and independence of walking time between individual trips; the validity of these assumptions may depend on when people travel. When travelling during the weekend, for example, people may be more inclined to undertake an activity during transfers, since they will not be in the weekday pattern of rushing to and from work. In addition, during the morning rush hour, travellers may focus more on getting to work on time than on other activities, reducing the share of people undertaking other activities during a transfer. If this is the case, then the model could include a hierarchical layer for the day of the week, and for the time of day. Having this layer would also give the public transit agencies the opportunity of having time-variable transfer times during the week and day, giving passengers better information when planning their journeys. The variation in travel behaviour will affect the number of people travelling across the day, with (by definition) more people in the peak periods. Should stations be overcrowded (not an issue for the Danish station used in the study), the assumption of independence between trips becomes invalid. When overcrowding occurs, the direct walking time at the station will likely be affected, limiting the use of the method to stations without overcrowding: overcrowding makes transfers between transportation modes take longer. If most travel occurs during the overcrowding periods, and the public transit agency uses the same scheduled transfer time for the whole day, then the inferred transfer times may still be helpful, albeit presumably too low. However, if the inferred transfer times are lower higher than the scheduled transfer time, then the scheduled transfer time should be increased. However, a better approach would be to adjust the model to handle overcrowding. In addition to incorporating overcrowding and time-varying behaviour, an interesting further extension would be to handle bus-to-bus and train-to-train transfers, as this would give transit agencies possibilities to improve the scheduling even further. However, both of these suggestions pose additional challenges. For the train-to-train transfers, the ridden train is unknown, making it difficult to assess the traveller's arrival time. And, in the case of buses, the travellers can only tap in once the bus has arrived, making it difficult to split the transfer time into walking time and waiting time.

5.1.3 Exploratory study in chapter 4

The approach for inferring the *area of interest* in chapter 4 culminated in the proposed models having large assumptions, in which the resulting ratios indicating the *area of interest* may be explained by other behavioural factors. However, as debated in the discussion section of the chapter 4, the assumptions could be relaxed by incorporating these behavioural factors into the proposed models, which presumably would give a more realistic model. Further research is presented in section 4.3.1 relating to the *areas of interest*. However, there is also a different direction for using AFC data with the online journey search data. In a preliminary investigation of data sources, a recurrent neural network (RNN) was used to see if it could improve the results of Roosmalen (2019) and Wang (2020) to predict the travel demand for public transit at the route level. The RNN model showed the same results as Roosmalen (2019) and Wang (2020). There was clear sign of seasonality effects and, due to there being only a single year of data to examine, the model was not able to learn the seasonal pattern effectively. If several years of data were available, it would presumably improve the results, which would be an interesting and useful outcome for public transit agencies. This research path would be simpler for further research, since the results can be tested with relative ease.

5.1.4 A data cleaning framework

Even though this thesis has contributed to the existing literature with new knowledge and methods on missing information of the AFC system, further research is needed to improve the attractiveness and efficiency of public transportation.

An impactful future project is the creation of a *data cleaning framework* (Robinson et al., 2014) for pre-processing and cleaning the complete transportation network data set. The aim of a *data cleaning framework* can be achieved by building on the model from paper A by first making the improvement discussed in 5.1.3, then extending the model to several lines, and then including all lines of the network. Hereafter, the model could be expanded to handle additional missing information rooted in the AFC system, such as the direct and activity transfer times from paper B. Consolidating this into a single Bayesian model would give a complete picture of the public transportation network for further analysis. This framework may lower the workload of downstream analysis and make it simpler for researchers and practitioners to use. Building and inferring this large Bayesian model may be unrealistic in the short run. A more practicable approach would be to base the *data cleaning framework* on the five challenge categories—*design*, *software*, *hardware*, *input* and *user*—outlined in the introduction section 1.3, and consolidate the state-of-the-art-methods by building on these. Due to the thesis focussing only on the *design* challenge, a review of the current state of research for the four other challenges is needed to achieve a comprehensive and complete *data cleaning framework*. For the *design* challenge, the previous section presented further directions for research on the methods that this thesis contributes. Yet, there is still further research to be done on the *design* challenge. An interesting problem is that the AFC system contains only a subset of all trips performed in the network. In addition, for some smart cards, the complete trip chain is not collected, and some journeys may not be observed at all, as in the case of the Danish commuter and school cards. These limitations in the data restrict the public transit agencies from achieving a complete and accurate picture of the public transportation network. In the Danish case, additional information may be used to infer the missing trip chains and journeys. Cards such as the commuter and school cards are only valid in specific zones, limiting the possibilities for the paths used in the trip. Presuming that the cardholders use the cards to get and from work or school, thus using the same path every time, it may be possible with data from conductor controls to infer the missing trip legs and journeys.

The *data cleaning framework* could be carried out as an open-source project, where the methods are implemented in a standardised way (Lawson et al., 2019). Standardising the method would make it easier for public transit agencies to adopt and use state-of-the-art research, thereby improving the public transportation network for the general public. In addition, having an open-source project would also presumably encourage collaboration between practitioners and researchers. However, the framework should be managed by a company in the industry, a research centre, or both, to ensure continued development.

Appendix A

Some of the different card types in the Danish AFC system.

The following briefly describes some of the different types of smart cards and selected rules associated with them. The rule for open-system cards can vary depending on the agency responsible for the region.

- *Anonymous* The anonymous card is not associated with any specific person and does not have any special benefits associated with it. The user is required to tap in at every change of transportation mode, and to tap out at the end of the trip.

- *Personal* The personal card is associated with a specific person and is only allowed to be used by this person. This card is granted, in some areas, a discount depending on how much the card owner uses the card, and there can be further discounts for students, senior citizens or groups etc. The user is required to tap in every time the user changes transportation mode and to tap out at the end of the trip.

- *Flex* This card is registered to a specific person, but can be used by anyone. The card has some, but not all, of the discount benefits of the Personal card.
- *Company/Business* The card is like the Flex card, but instead of being registered to a person, it is registered to a company.
- *Commute* The Commute card is person-specific, and valid only in a given area for a given monthly price. The card cannot be used outside its assigned area. The commuter is only required to tap in when activating a travel period, and when using buses.
- *Combo-Commute* The Combo-Commute card combines features of the personal and commute cards. It can be used outside of its assigned commuting area, but the user is required to tap in and out during every trip, even when in the assigned area.
- *Youth* This card has the same setup as the commute card, but is for people aged 16–19 or students. It is valid only in zones between the home address and the place of education.
- *School* The School card has the same setup as the Commute card, but is for people in high schools. It is valid only in zones between the home address and the place of education.

Appendix B

Identifying areas of interest.

Table B.1: Demographics of the municipalities and the probability of taking a trip and searching - Continues on next page.

				$P(Trip^{\text{From}})$		$P(Trip^{\text{To}})$		$P(Seaches^{\text{From}})$		$P(Seaches^{\text{To}})$	
Municipality	Pop. Rank	Population	Employed	Rank	$P(-)$	Rank	$P(-)$	Rank	$P(-)$	Rank	$P(-)$
København	1	613288	404947	0	32.52%	0	32.33%	0	22.85%	0	22.03%
Aarhus	2	340421	202156	1	6.07%	1	6.06%	1	7.9%	1	8.04%
Aalborg	3	213558	112794	2	5.95%	2	5.94%	3	3.82%	3	3.88%
Odense	4	202348	105665	4	2.94%	4	2.98%	2	5.41%	2	6.47%
Esbjerg	5	116032	60373	12	1.4%	12	1.41%	8	1.63%	8	1.71%
Vejle	6	114140	56178	9	1.55%	9	1.54%	7	1.69%	7	1.72%
Frederiksberg	7	104410	42368	3	4.44%	3	4.45%	6	1.69%	6	1.87%
Randers	8	98265	41674	48	0.41%	46	0.41%	27	0.86%	26	0.84%
Viborg	9	96883	50683	52	0.34%	52	0.34%	30	0.74%	28	0.76%
Kolding	10	92515	54183	16	1.17%	16	1.17%	12	1.46%	13	1.47%
Silkeborg	11	92024	40789	44	0.43%	44	0.43%	31	0.71%	30	0.73%
Horsens	12	89598	44226	59	0.29%	56	0.3%	23	0.96%	24	0.91%
Herning	13	88733	47488	58	0.29%	57	0.3%	26	0.88%	23	0.98%
Roskilde	14	87382	43156	8	1.94%	8	1.95%	4	3.25%	4	3.71%
Næstved	15	82938	31284	21	0.92%	21	0.92%	13	1.41%	15	1.36%
Slagelse	16	78968	34477	22	0.88%	22	0.88%	9	1.59%	12	1.5%
Gentofte	17	75803	39873	7	1.95%	7	2.0%	14	1.38%	11	1.51%
Sønderborg	18	74650	33553	23	0.83%	23	0.83%	45	0.57%	36	0.6%
Holbæk	19	70983	28308	20	0.94%	20	0.93%	16	1.29%	18	1.22%
Gladsaxe	20	69484	43995	10	1.52%	10	1.48%	18	1.23%	20	1.04%
Hjørring	21	65257	30409	39	0.45%	39	0.45%	48	0.56%	41	0.58%
Helsingør	22	62686	23014	19	0.95%	19	0.95%	19	1.17%	19	1.11%
Guldborgsund	23	61219	23473	36	0.5%	36	0.5%	32	0.71%	33	0.68%
Skanderborg	24	61158	27745	47	0.41%	48	0.41%	33	0.7%	32	0.68%
Køge	25	60356	29277	18	1.05%	18	1.05%	15	1.38%	14	1.38%
Frederikshavn	26	60140	27346	50	0.38%	50	0.38%	50	0.51%	45	0.54%

Table B.2: Demographics of the municipalities and the probability of taking a trip and searching - Continues on next page.

				$P(Trip^{\text{From}})$		$P(Trip^{\text{To}})$		$P(Seaches^{\text{From}})$		$P(Seaches^{\text{To}})$	
Municipality	Pop. Rank	Population	Employed	Rank	$P(-)$	Rank	$P(-)$	Rank	$P(-)$	Rank	$P(-)$
Aabenraa	27	59089	29413	29	0.64%	29	0.64%	54	0.46%	50	0.48%
Svendborg	28	58698	23462	32	0.57%	31	0.57%	34	0.68%	29	0.74%
Holstebro	29	58418	31198	82	0.13%	82	0.13%	52	0.5%	46	0.53%
Ringkøbing-Skjern	30	57005	30261	84	0.11%	84	0.11%	66	0.34%	64	0.33%
Rudersdal	31	55989	27194	14	1.2%	14	1.21%	20	1.15%	21	1.02%
Haderslev	32	55963	23461	31	0.57%	32	0.57%	64	0.35%	61	0.35%
Lyngby-Taarbæk	33	55472	35854	5	2.09%	5	2.12%	11	1.48%	10	1.62%
Hvidovre	34	53282	29801	11	1.44%	11	1.42%	22	1.13%	22	0.98%
Faaborg-Midtfyn	35	51536	19225	51	0.36%	51	0.36%	55	0.45%	58	0.38%
Fredericia	36	51326	27917	41	0.45%	41	0.45%	25	0.89%	27	0.77%
Hillerød	37	50650	30723	17	1.09%	17	1.09%	21	1.15%	16	1.31%
Høje-Taastrup	38	50596	38781	15	1.19%	15	1.2%	5	1.88%	9	1.66%
Varde	39	50301	22201	57	0.3%	60	0.29%	65	0.34%	62	0.34%
Greve	40	49974	18679	25	0.77%	25	0.78%	29	0.76%	34	0.65%
Kalundborg	41	48982	19311	46	0.41%	47	0.41%	59	0.4%	55	0.39%
Ballerup	42	48295	43191	13	1.28%	13	1.28%	17	1.27%	17	1.25%
Favrskov	43	48271	19470	68	0.23%	68	0.23%	62	0.37%	60	0.35%
Hedensted	44	46616	21263	85	0.11%	85	0.11%	81	0.22%	80	0.21%
Skive	45	46599	21797	83	0.12%	83	0.12%	76	0.25%	74	0.26%
Vordingborg	46	46087	16425	43	0.43%	43	0.43%	46	0.57%	39	0.58%
Frederikssund	47	45189	16477	38	0.46%	38	0.46%	37	0.65%	40	0.58%
Thisted	48	43716	22112	72	0.2%	71	0.2%	68	0.29%	68	0.29%
Tårnby	49	43063	26916	6	2.03%	6	2.08%	10	1.57%	5	2.47%
Egedal	50	43000	12153	40	0.45%	40	0.45%	36	0.65%	47	0.52%
Vejen	51	42844	20563	61	0.28%	61	0.28%	61	0.37%	66	0.32%
Syddjurs	52	42468	14667	75	0.18%	75	0.18%	71	0.27%	70	0.28%
Mariagerfjord	53	42125	20838	74	0.19%	74	0.19%	57	0.42%	57	0.38%
Lolland	54	41982	16320	55	0.31%	54	0.31%	74	0.26%	72	0.27%
Assens	55	41328	15242	67	0.24%	67	0.23%	56	0.42%	59	0.36%
Gribskov	56	41217	13108	49	0.4%	49	0.4%	63	0.37%	56	0.38%
Ikast-Brande	57	41191	22913	87	0.08%	87	0.08%	75	0.26%	77	0.24%
Furesø	58	40911	14110	26	0.76%	27	0.76%	35	0.68%	38	0.59%
Fredensborg	59	40779	13043	34	0.52%	35	0.51%	43	0.59%	48	0.52%
Rødovre	60	39343	17363	30	0.61%	30	0.64%	51	0.5%	51	0.45%
Jammerbugt	61	38638	14778	66	0.24%	66	0.24%	77	0.25%	75	0.25%

Table B.3: Demographics of the municipalities and the probability of taking a trip and searching.

				$P(T^{\text{From}})$		$P(T^{\text{To}})$		$P(S^{\text{From}})$		$P(S^{\text{To}})$	
Municipality	Pop. Rank	Population	Employed	Rank	$P(-)$	Rank	$P(-)$	Rank	$P(-)$	Rank	$P(-)$
Middelfart	62	38210	17869	78	0.18%	77	0.17%	42	0.6%	42	0.57%
Norddjurs	63	38197	15797	89	0.05%	89	0.06%	89	0.11%	89	0.13%
Tønder	64	37777	16496	60	0.29%	59	0.29%	82	0.21%	79	0.21%
Vesthimmerlands	65	37277	17457	80	0.15%	80	0.15%	83	0.21%	82	0.21%
Brønderslev	66	36289	13404	70	0.21%	70	0.21%	69	0.27%	73	0.26%
Faxe	67	36139	12596	65	0.26%	65	0.26%	67	0.31%	69	0.28%
Brøndby	68	35538	24982	28	0.64%	28	0.65%	40	0.62%	44	0.55%
Ringsted	69	34473	16871	37	0.48%	37	0.48%	24	0.89%	25	0.87%
Odsherred	70	33083	11309	54	0.31%	53	0.31%	73	0.26%	71	0.28%
Nyborg	71	32032	11300	62	0.28%	62	0.27%	28	0.81%	31	0.73%
Halsnæs	72	31168	9197	64	0.27%	64	0.27%	70	0.27%	76	0.25%
Rebild	73	29827	11657	76	0.18%	76	0.18%	80	0.23%	81	0.21%
Sorø	74	29669	11970	63	0.27%	63	0.27%	49	0.52%	52	0.45%
Nordfyns	75	29516	10098	77	0.18%	78	0.17%	87	0.15%	87	0.14%
Herlev	76	28572	23183	27	0.75%	26	0.77%	39	0.62%	35	0.63%
Albertslund	77	27743	20788	33	0.56%	33	0.57%	44	0.59%	49	0.49%
Lejre	78	27544	8178	56	0.3%	58	0.3%	47	0.56%	53	0.45%
Billund	79	26482	18768	73	0.19%	72	0.2%	78	0.24%	65	0.32%
Allerød	80	25235	14534	42	0.43%	42	0.44%	53	0.46%	54	0.42%
Hørsholm	81	25028	9646	35	0.52%	34	0.52%	38	0.63%	43	0.55%
Kerteminde	82	23756	9643	71	0.21%	73	0.2%	72	0.26%	83	0.2%
Ishøj	83	22988	9417	45	0.42%	45	0.41%	58	0.4%	63	0.33%
Stevns	84	22727	6184	81	0.15%	81	0.15%	85	0.18%	86	0.15%
Glostrup	85	22663	21889	24	0.77%	24	0.79%	41	0.61%	37	0.59%
Odder	86	22626	7884	86	0.09%	86	0.09%	88	0.12%	88	0.13%
Solrød	87	22518	6319	53	0.31%	55	0.3%	60	0.39%	67	0.3%
Struer	88	21270	8234	88	0.07%	88	0.07%	79	0.24%	78	0.24%
Morsø	89	20514	9457	92	0.02%	92	0.02%	92	0.06%	91	0.06%
Lemvig	90	20133	9329	91	0.02%	91	0.02%	90	0.06%	90	0.07%
Vallensbæk	91	16280	5754	69	0.21%	69	0.21%	84	0.2%	84	0.16%
Dragør	92	14272	3207	79	0.16%	79	0.17%	86	0.16%	85	0.15%
Langeland	93	12641	4514	90	0.04%	90	0.04%	91	0.06%	92	0.05%

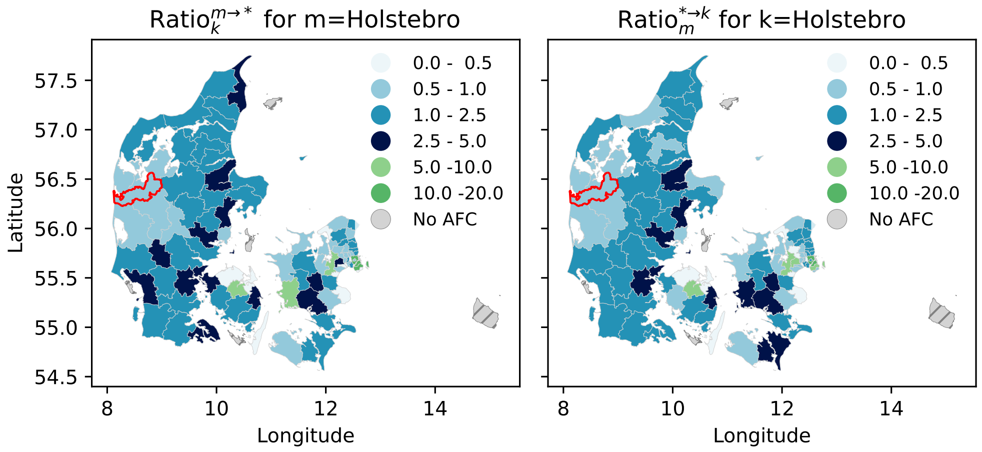
B.1 Identifying areas of interest.

Table B.4: Top Ratio$_k^{m \to *}$ with the four largest municipalities and Holstebro - Condition from the municipality Holstebro.

To	Population	Distance	Trips	Searches	$\widetilde{pi}_k^{m->*}$	$\widetilde{\lambda}_k^{m->*}$	Ratio$_k^{m->*}$	Rank
Tårnby	43,063	265km	78	23,030	0.09%	1.37%	14.97	1
København	613,288	257km	1,208	152,989	0.67%	9.08%	13.48	2
Roskilde	87,382	232km	89	11,333	0.10%	0.68%	6.97	3
Odense	202,348	155km	688	42,794	0.41%	2.54%	6.27	4
Slagelse	78,968	206km	126	10,068	0.12%	0.60%	5.18	5
Høje-Taastrup	50,596	239km	247	14,218	0.18%	0.85%	4.75	6
Ringsted	34,473	225km	66	5,977	0.09%	0.36%	4.22	7
Aarhus	340,421	99km	5,462	145,326	2.86%	8.63%	3.01	14
Aalborg	213,558	110km	4,057	48,235	2.14%	2.87%	1.34	42
Holstebro	58,418	0km	75,993	370,073	39.18%	21.96%	0.56	83

Table B.5: Top Ratio$_m^{* \to k}$ with the four largest municipalities and Holstebro - Condition to municipality Holstebro.

To	Population	Distance	Trips	Searches	$\widetilde{\pi}_k^{*->k}$	$\widetilde{\lambda}_k^{*->k}$	Ratio$_k^{*->k}$	Rank
København	613,288	257km	1,324	156,073	0.73%	8.72%	11.92	1
Tårnby	43,063	265km	138	21,165	0.12%	1.19%	9.71	2
Roskilde	87,382	232km	84	11,825	0.09%	0.67%	7.04	3
Odense	202,348	155km	715	49,140	0.42%	2.75%	6.57	4
Høje-Taastrup	50,596	239km	263	19,426	0.19%	1.09%	5.85	5
Ringsted	34,473	225km	61	6,522	0.08%	0.37%	4.47	6
Slagelse	78,968	206km	165	10,397	0.14%	0.59%	4.30	7
Aarhus	340,421	99km	5,345	150,000	2.80%	8.38%	3.00	14
Aalborg	213,558	110km	4,400	44,685	2.31%	2.50%	1.08	56
Holstebro	58,418	0km	75,993	370,073	39.10%	20.67%	0.53	83



(a) From Holstebro.

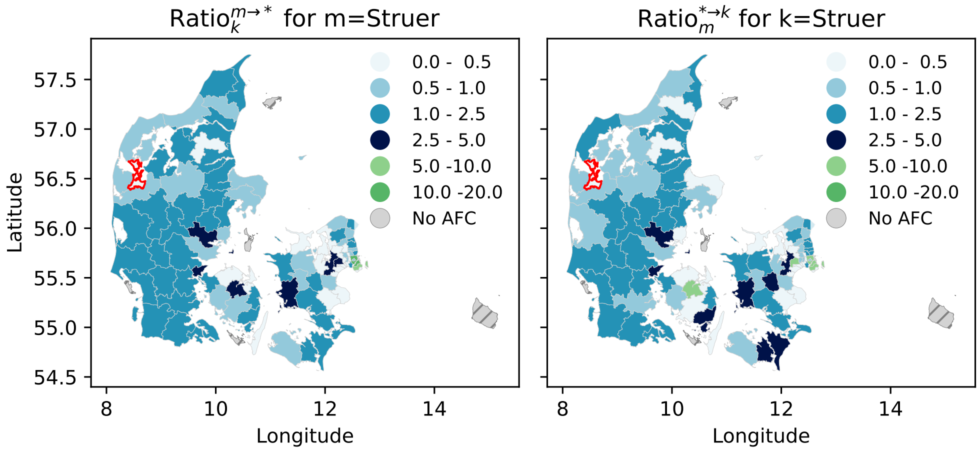
Figure B.1: Ratio between the conditional probability of travelling and searching in relation of Holstebro.

Table B.6: Top $\text{Ratio}_k^{m\to *}$ with the four largest municipalities and Struer - Condition from the municipality Struer.

To	Population	Distance	Trips	Searches	$\widetilde{pi}_k^{m->*}$	$\widetilde{\lambda}_k^{m->*}$	$\text{Ratio}_k^{m->*}$	Rank
København	613,288	265km	756	71,593	0.85%	8.65%	10.18	1
Tårnby	43,063	273km	26	10,066	0.12%	1.23%	9.81	2
Odense	202,348	170km	366	18,604	0.46%	2.26%	4.88	3
Høje-Taastrup	50,596	248km	138	9,087	0.24%	1.11%	4.69	4
Roskilde	87,382	241km	53	5,213	0.15%	0.64%	4.22	5
Slagelse	78,968	218km	76	5,673	0.17%	0.70%	3.99	6
Fredericia	51,326	126km	334	10,462	0.43%	1.27%	2.96	7
Aarhus	340,421	107km	4,413	82,020	4.48%	9.90%	2.21	14
Aalborg	213,558	102km	808	15,271	0.90%	1.85%	2.06	16
Struer	21,270	0km	29,182	113,789	29.04%	13.74%	0.47	73

Table B.7: Top $\text{Ratio}_m^{*\to k}$ with the four largest municipalities and Struer - Condition to municipality Struer.

To	Population	Distance	Trips	Searches	$\widetilde{\pi}_k^{*->k}$	$\widetilde{\lambda}_k^{*->k}$	$\text{Ratio}_k^{*->k}$	Rank
København	613,288	265km	810	70,479	0.91%	8.55%	9.43	1
Tårnby	43,063	273km	77	10,458	0.18%	1.28%	7.26	2
Odense	202,348	170km	353	20,517	0.45%	2.50%	5.54	3
Høje-Taastrup	50,596	248km	163	10,814	0.26%	1.32%	5.05	4
Fredericia	51,326	126km	333	16,200	0.43%	1.97%	4.58	5
Roskilde	87,382	241km	67	5,498	0.17%	0.68%	4.08	6
Guldborgsund	61,219	282km	16	3,305	0.12%	0.41%	3.57	7
Aarhus	340,421	107km	4,178	87,279	4.26%	10.58%	2.48	12
Aalborg	213,558	102km	901	13,560	1.00%	1.65%	1.66	20
Struer	21,270	0km	29,182	113,789	29.15%	13.79%	0.47	76



(a) From Struer.

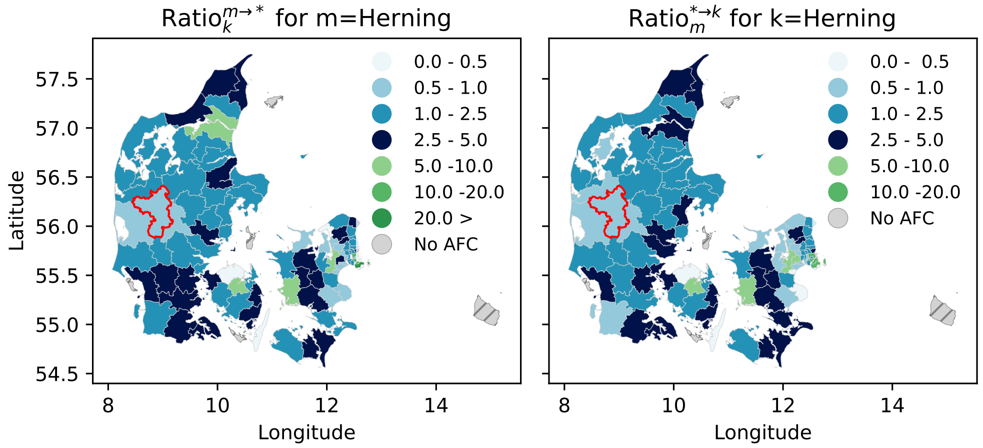
Figure B.2: Ratio between the conditional probability of travelling and searching in relation of Struer.

Table B.8: Top Ratio$_k^{m\to *}$ with the four largest municipalities and Herning - Condition from the municipality Herning.

To	Population	Distance	Trips	Searches	$\widetilde{pi}_k^{m->*}$	$\widetilde{\lambda}_k^{m->*}$	Ratio$_k^{m->*}$	Rank
Tårnby	43,063	242km	119	38,728	0.05%	1.30%	24.88	1
København	613,288	235km	2,818	224,268	0.70%	7.51%	10.79	2
Odense	202,348	127km	1,375	71,645	0.35%	2.40%	6.83	3
Slagelse	78,968	179km	185	13,435	0.07%	0.45%	6.67	4
Aalborg	213,558	116km	915	43,354	0.24%	1.45%	6.01	5
Roskilde	87,382	209km	248	14,096	0.08%	0.48%	5.73	6
Høje-Taastrup	50,596	217km	454	19,083	0.13%	0.64%	4.86	7
Svendborg	58,698	160km	108	6,965	0.05%	0.24%	4.77	8
Aarhus	340,421	79km	18,502	264,453	4.44%	8.86%	2.00	40
Herning	88,733	0km	242,573	988,873	57.88%	33.11%	0.57	85

Table B.9: Top Ratio$_m^{*\to k}$ with the four largest municipalities and Herning - Condition to municipality Herning.

To	Population	Distance	Trips	Searches	$\widetilde{\pi}_k^{*->k}$	$\widetilde{\lambda}_k^{*->k}$	Ratio$_k^{*->k}$	Rank
Tårnby	43,063	242km	219	36,108	0.08%	1.09%	14.50	1
København	613,288	235km	2,778	226,895	0.68%	6.85%	10.07	2
Odense	202,348	127km	1,255	78,946	0.32%	2.39%	7.45	3
Slagelse	78,968	179km	198	16,021	0.07%	0.49%	6.91	4
Høje-Taastrup	50,596	217km	409	22,313	0.12%	0.68%	5.62	5
Roskilde	87,382	209km	256	15,191	0.08%	0.46%	5.49	6
Ringsted	34,473	200km	142	9,243	0.06%	0.28%	4.93	7
Aalborg	213,558	116km	1,034	43,641	0.27%	1.32%	4.93	8
Aarhus	340,421	79km	19,091	500,470	4.53%	15.10%	3.33	20
Herning	88,733	0km	242,573	988,873	57.33%	29.84%	0.52	87



(a) From Herning.

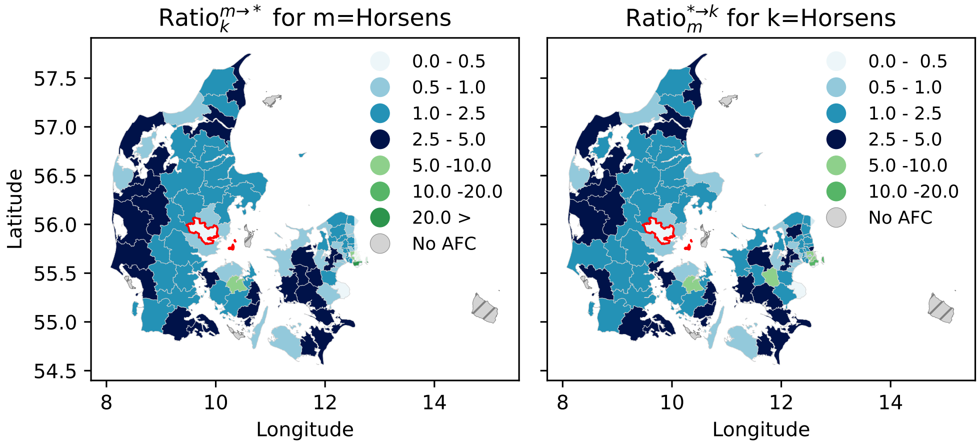
Figure B.3: Ratio between the conditional probability of travelling and searching in relation of Herning.

Table B.10: Top $\text{Ratio}_k^{m \to *}$ with the four largest municipalities and Horsens - Condition from the municipality Horsens.

To	Population	Distance	Trips	Searches	$\widetilde{pi}_k^{m->*}$	$\widetilde{\lambda}_k^{m->*}$	$\text{Ratio}_k^{m->*}$	Rank
Tårnby	43,063	182km	309	82,552	0.10%	2.53%	25.85	1
København	613,288	175km	4,754	272,102	1.16%	8.32%	7.17	2
Odense	202,348	71km	2,953	130,214	0.73%	3.98%	5.46	3
Roskilde	87,382	149km	530	22,323	0.15%	0.69%	4.55	4
Holstebro	58,418	88km	218	10,811	0.08%	0.33%	4.39	5
Slagelse	78,968	118km	394	16,743	0.12%	0.51%	4.36	6
Ringsted	34,473	138km	293	13,133	0.09%	0.40%	4.31	7
Aalborg	213,558	121km	2,681	68,468	0.66%	2.10%	3.15	18
Aarhus	340,421	35km	64,814	838,609	15.52%	25.64%	1.65	45
Horsens	89,598	0km	225,062	707,310	53.82%	21.63%	0.40	90

Table B.11: Top $\text{Ratio}_m^{* \to k}$ with the four largest municipalities and Horsens - Condition to municipality Horsens.

To	Population	Distance	Trips	Searches	$\widetilde{\pi}_k^{*->k}$	$\widetilde{\lambda}_k^{*->k}$	$\text{Ratio}_k^{*->k}$	Rank
Tårnby	43,063	182km	458	65,516	0.13%	2.13%	16.18	1
København	613,288	175km	4,712	249,777	1.14%	8.12%	7.14	2
Odense	202,348	71km	2,930	133,872	0.72%	4.35%	6.08	3
Ringsted	34,473	138km	310	14,924	0.10%	0.49%	5.04	4
Roskilde	87,382	149km	538	22,453	0.15%	0.73%	4.86	5
Holstebro	58,418	88km	203	10,359	0.07%	0.34%	4.75	6
Slagelse	78,968	118km	411	16,333	0.12%	0.53%	4.42	7
Aalborg	213,558	121km	2,601	62,638	0.64%	2.04%	3.20	17
Aarhus	340,421	35km	67,082	733,057	15.86%	23.81%	1.50	49
Horsens	89,598	0km	225,062	707,310	53.16%	22.98%	0.43	90



(a) From Horsens.

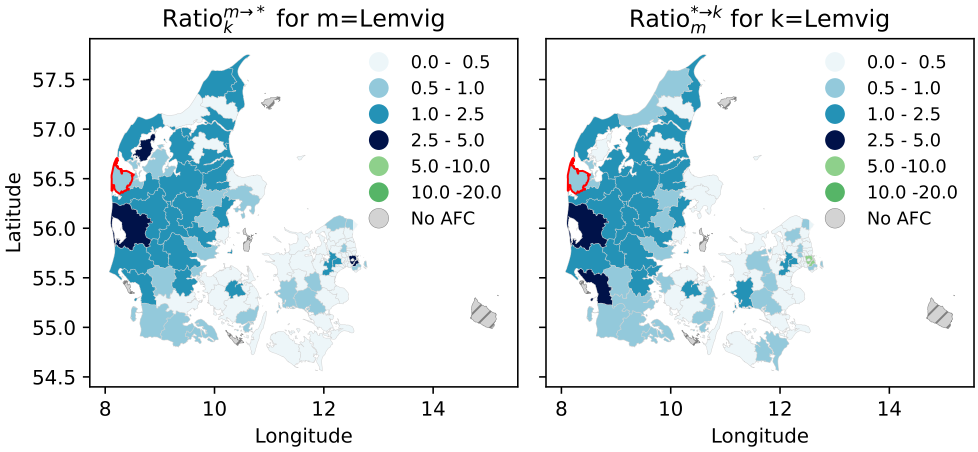
Figure B.4: Ratio between the conditional probability of travelling and searching in relation of Horsens.

Table B.12: Top Ratio$_k^{m\rightarrow *}$ with the four largest municipalities and Lemvig - Condition from the municipality Lemvig.

To	Population	Distance	Trips	Searches	$\widetilde{pi}_k^{m->*}$	$\widetilde{\lambda}_k^{m->*}$	Ratio$_k^{m->*}$	Rank
København	613,288	280km	112	4,752	0.50%	2.14%	4.25	1
Morsø	20,514	43km	20	1,961	0.28%	0.91%	3.19	2
Ringkøbing-Skjern	57,005	54km	255	5,472	0.84%	2.46%	2.92	3
Aarhus	340,421	121km	390	5,990	1.16%	2.69%	2.31	4
Aalborg	213,558	117km	141	2,551	0.57%	1.17%	2.04	5
Thisted	43,716	54km	123	2,097	0.53%	0.97%	1.83	6
Esbjerg	116,032	122km	84	1,702	0.44%	0.79%	1.82	7
Herning	88,733	52km	772	8,391	2.07%	3.75%	1.81	8
Odense	202,348	179km	41	1,232	0.33%	0.59%	1.76	9
Lemvig	20,133	0km	17,191	78,825	41.04%	34.82%	0.85	32

Table B.13: Top Ratio$_m^{*\rightarrow k}$ with the four largest municipalities and Lemvig - Condition to municipality Lemvig.

To	Population	Distance	Trips	Searches	$\widetilde{\pi}_k^{*->k}$	$\widetilde{\lambda}_k^{*->k}$	Ratio$_k^{*->k}$	Rank
København	613,288	280km	142	7,263	0.57%	3.18%	5.62	1
Ringkøbing-Skjern	57,005	54km	242	6,385	0.80%	2.80%	3.50	2
Esbjerg	116,032	122km	74	2,313	0.41%	1.04%	2.56	3
Aarhus	340,421	121km	454	7,317	1.29%	3.20%	2.47	4
Kolding	92,515	136km	36	1,475	0.32%	0.68%	2.14	5
Odense	202,348	179km	53	1,622	0.36%	0.74%	2.08	6
Aalborg	213,558	117km	188	3,111	0.67%	1.39%	2.06	7
Herning	88,733	52km	709	7,300	1.89%	3.19%	1.69	8
Thisted	43,716	54km	134	1,980	0.55%	0.90%	1.64	9
Lemvig	20,133	0km	17,191	78,825	40.41%	34.07%	0.84	31



(a) From Lemvig.

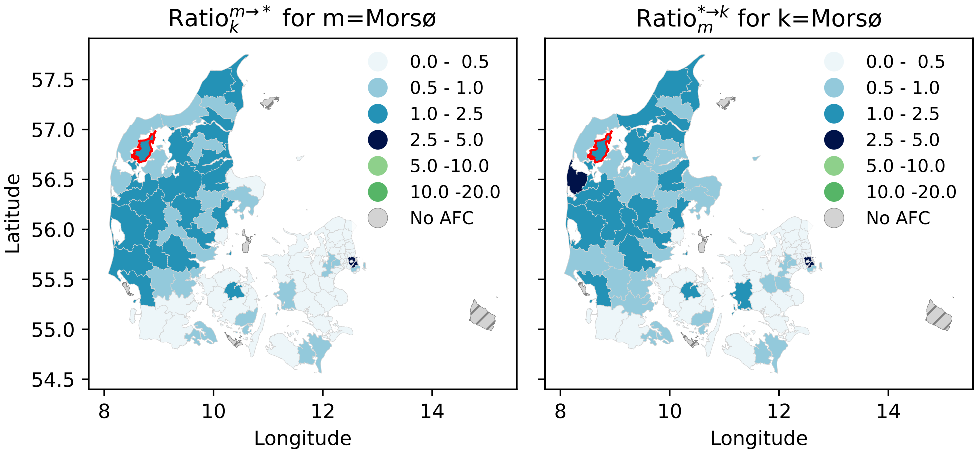
Figure B.5: Ratio between the conditional probability of travelling and searching in relation of Lemvig.

Table B.14: Top Ratio$_k^{m\rightarrow *}$ with the four largest municipalities and Morsø - Condition from the municipality Morsø.

To	Population	Distance	Trips	Searches	$\widetilde{pi}_k^{m->*}$	$\widetilde{\lambda}_k^{m->*}$	Ratio$_k^{m->*}$	Rank
København	613,288	267km	37	2,606	0.34%	1.35%	3.93	1
Varde	50,301	126km	5	1,037	0.26%	0.57%	2.16	2
Frederikshavn	60,140	123km	57	1,441	0.39%	0.77%	1.95	3
Aarhus	340,421	113km	649	7,003	1.87%	3.54%	1.89	4
Aalborg	213,558	79km	1,440	13,549	3.85%	6.79%	1.77	5
Hjørring	65,257	108km	65	1,350	0.41%	0.72%	1.75	6
Vesthimmerlands	37,277	40km	76	1,300	0.44%	0.70%	1.58	7
Holstebro	58,418	48km	287	2,956	0.97%	1.52%	1.57	8
Morsø	20,514	0km	8,788	67,585	22.21%	33.69%	1.52	11
Odense	202,348	187km	31	841	0.33%	0.47%	1.43	14

Table B.15: Top Ratio$_m^{*\rightarrow k}$ with the four largest municipalities and Morsø - Condition to municipality Morsø.

To	Population	Distance	Trips	Searches	$\widetilde{\pi}_k^{*->k}$	$\widetilde{\lambda}_k^{*->k}$	Ratio$_k^{*->k}$	Rank
København	613,288	267km	65	3,453	0.41%	1.78%	4.36	1
Lemvig	20,133	43km	20	1,961	0.30%	1.03%	3.48	2
Odense	202,348	187km	32	1,081	0.33%	0.59%	1.81	3
Holstebro	58,418	48km	280	3,202	0.94%	1.66%	1.76	4
Randers	98,265	88km	67	1,295	0.41%	0.70%	1.69	5
Hjørring	65,257	108km	59	1,209	0.39%	0.66%	1.67	6
Aarhus	340,421	113km	844	7,628	2.34%	3.88%	1.66	7
Vesthimmerlands	37,277	40km	61	1,208	0.40%	0.66%	1.65	8
Aalborg	213,558	79km	1,483	12,167	3.92%	6.16%	1.57	11
Morsø	20,514	0km	8,788	67,585	22.00%	33.96%	1.54	13



(a) From Morsø.

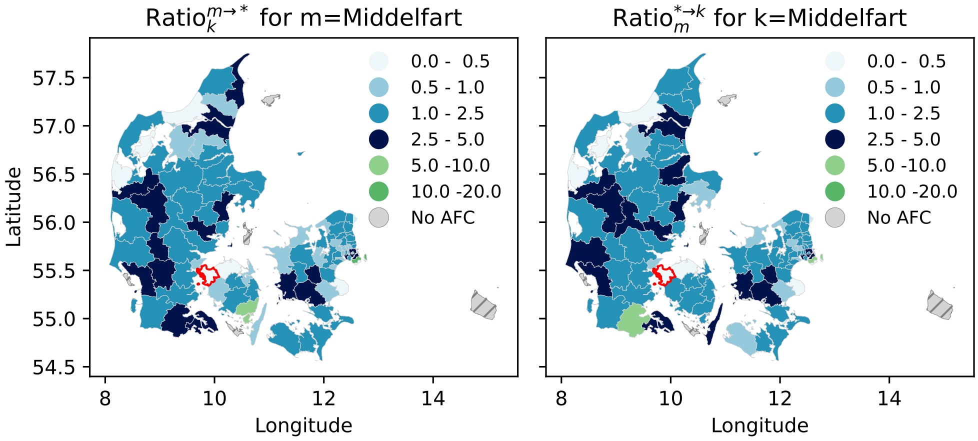
Figure B.6: Ratio between the conditional probability of travelling and searching in relation of Morsø.

Table B.16: Top Ratio$_k^{m \to *}$ with the four largest municipalities and Middelfart - Condition from the municipality Middelfart.

To	Population	Distance	Trips	Searches	$\widetilde{pi}_k^{m->*}$	$\widetilde{\lambda}_k^{m->*}$	Ratio$_k^{m->*}$	Rank
Tårnby	43,063	174km	383	52,959	0.19%	2.59%	13.60	1
Svendborg	58,698	60km	960	67,324	0.42%	3.29%	7.87	2
København	613,288	170km	6,124	236,363	2.46%	11.55%	4.70	3
Herning	88,733	101km	189	10,724	0.11%	0.53%	4.64	4
Esbjerg	116,032	76km	1,211	46,984	0.52%	2.30%	4.45	5
Aalborg	213,558	174km	422	18,314	0.21%	0.90%	4.37	6
Sønderborg	74,650	56km	262	11,986	0.14%	0.59%	4.13	7
Aarhus	340,421	81km	3,857	119,334	1.56%	5.83%	3.74	8
Odense	202,348	32km	35,263	403,192	13.96%	19.70%	1.41	48
Middelfart	38,210	0km	134,271	379,266	53.03%	18.53%	0.35	89

Table B.17: Top Ratio$_m^{* \to k}$ with the four largest municipalities and Middelfart - Condition to municipality Middelfart.

To	Population	Distance	Trips	Searches	$\widetilde{\pi}_k^{*->k}$	$\widetilde{\lambda}_k^{*->k}$	Ratio$_k^{*->k}$	Rank
Tårnby	43,063	174km	618	39,665	0.28%	2.04%	7.16	1
Aabenraa	59,089	64km	382	18,757	0.19%	0.97%	5.06	2
København	613,288	170km	6,270	221,928	2.53%	11.39%	4.51	3
Aalborg	213,558	174km	435	17,019	0.21%	0.88%	4.14	4
Herning	88,733	101km	218	9,839	0.13%	0.51%	4.04	5
Aarhus	340,421	81km	3,752	116,492	1.53%	5.98%	3.92	6
Esbjerg	116,032	76km	1,172	37,456	0.50%	1.93%	3.82	7
Ringsted	34,473	122km	706	22,685	0.32%	1.17%	3.66	8
Odense	202,348	32km	33,700	382,646	13.41%	19.64%	1.46	49
Middelfart	38,210	0km	134,271	379,266	53.31%	19.47%	0.37	88



(a) From Middelfart.

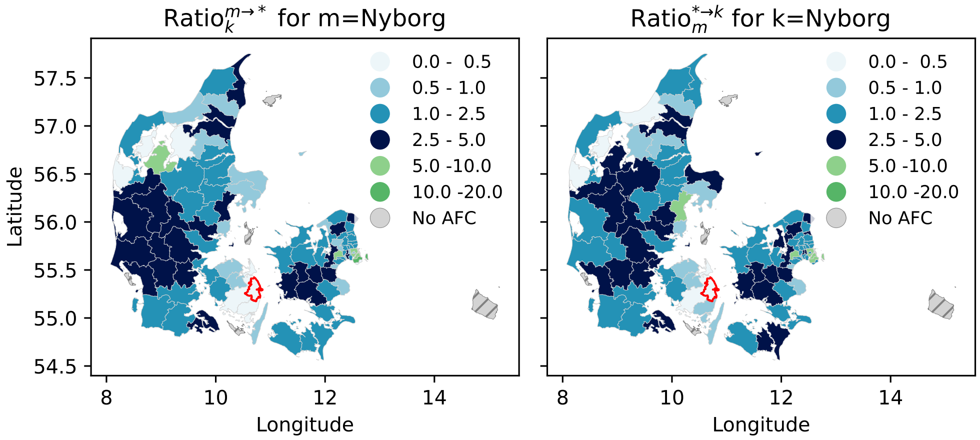
Figure B.7: Ratio between the conditional probability of travelling and searching in relation of Middelfart.

Table B.18: Top Ratio$_k^{m\rightarrow *}$ with the four largest municipalities and Nyborg - Condition from the municipality Nyborg.

To	Population	Distance	Trips	Searches	$\widetilde{pi}_k^{m->*}$	$\widetilde{\lambda}_k^{m->*}$	Ratio$_k^{m->*}$	Rank
Tårnby	43,063	127km	1,927	205,606	0.52%	7.49%	14.47	1
Høje-Taastrup	50,596	105km	1,785	96,197	0.48%	3.51%	7.29	2
København	613,288	124km	19,801	778,192	5.08%	28.35%	5.58	3
Skive	46,599	184km	26	4,525	0.03%	0.17%	5.23	4
Slagelse	78,968	41km	4,498	155,286	1.17%	5.66%	4.82	5
Esbjerg	116,032	128km	503	20,185	0.15%	0.74%	4.80	6
Aalborg	213,558	197km	330	13,980	0.11%	0.51%	4.67	8
Aarhus	340,421	102km	1,505	48,704	0.41%	1.78%	4.34	10
Odense	202,348	23km	76,145	505,903	19.47%	18.43%	0.95	73
Nyborg	32,032	0km	195,116	241,461	49.86%	8.80%	0.18	93

Table B.19: Top Ratio$_m^{*\rightarrow k}$ with the four largest municipalities and Nyborg - Condition to municipality Nyborg.

To	Population	Distance	Trips	Searches	$\widetilde{\pi}_k^{*->k}$	$\widetilde{\lambda}_k^{*->k}$	Ratio$_k^{*->k}$	Rank
Tårnby	43,063	127km	2,985	158,712	0.80%	6.44%	8.09	1
Høje-Taastrup	50,596	105km	2,059	107,244	0.56%	4.35%	7.81	2
København	613,288	124km	19,752	691,510	5.12%	28.05%	5.47	3
Aarhus	340,421	102km	1,414	48,957	0.39%	1.99%	5.09	4
Slagelse	78,968	41km	4,552	145,517	1.20%	5.91%	4.92	5
Holstebro	58,418	177km	46	4,301	0.04%	0.18%	4.74	6
Esbjerg	116,032	128km	449	16,201	0.14%	0.66%	4.67	7
Aalborg	213,558	197km	479	13,517	0.15%	0.55%	3.70	16
Odense	202,348	23km	72,800	404,632	18.82%	16.42%	0.87	77
Nyborg	32,032	0km	195,116	241,461	50.39%	9.80%	0.19	93



(a) From Nyborg.

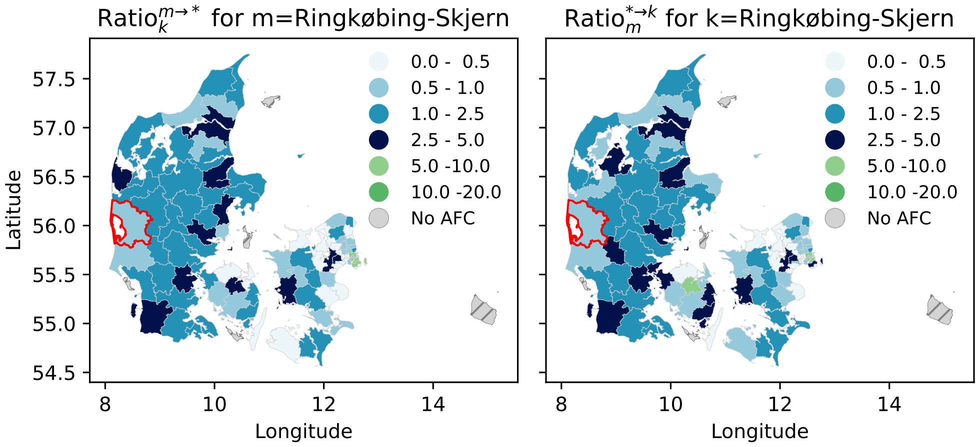
Figure B.8: Ratio between the conditional probability of travelling and searching in relation of Nyborg.

Table B.20: Top Ratio$_k^{m\to *}$ with the four largest municipalities and Ringkøbing-Skjern - Condition from the municipality Ringkøbing-Skjern.

To	Population	Distance	Trips	Searches	$\widetilde{pi}_k^{m->*}$	$\widetilde{\lambda}_k^{m->*}$	Ratio$_k^{m->*}$	Rank
København	613,288	258km	763	44,440	0.52%	3.82%	7.37	1
Tårnby	43,063	264km	31	5,973	0.08%	0.52%	6.62	2
Odense	202,348	138km	386	15,945	0.29%	1.38%	4.71	3
Høje-Taastrup	50,596	239km	98	5,529	0.12%	0.48%	4.06	4
Roskilde	87,382	231km	56	4,306	0.09%	0.38%	4.03	5
Aalborg	213,558	143km	328	10,300	0.26%	0.89%	3.47	6
Slagelse	78,968	195km	58	3,616	0.09%	0.32%	3.36	7
Aarhus	340,421	106km	5,182	120,313	3.17%	10.32%	3.25	8
Kolding	92,515	87km	397	11,130	0.30%	0.96%	3.23	9
Ringkøbing-Skjern	57,005	0km	90,355	411,494	54.32%	35.28%	0.65	63

Table B.21: Top Ratio$_m^{*\to k}$ with the four largest municipalities and Ringkøbing-Skjern - Condition to municipality Ringkøbing-Skjern.

To	Population	Distance	Trips	Searches	$\widetilde{\pi}_k^{*->k}$	$\widetilde{\lambda}_k^{*->k}$	Ratio$_k^{*->k}$	Rank
København	613,288	258km	734	49,183	0.50%	4.37%	8.69	1
Odense	202,348	138km	401	19,566	0.30%	1.74%	5.78	2
Høje-Taastrup	50,596	239km	104	6,687	0.12%	0.60%	4.89	3
Roskilde	87,382	231km	66	5,392	0.10%	0.49%	4.87	4
Tårnby	43,063	264km	38	4,379	0.08%	0.40%	4.78	5
Fredericia	51,326	91km	186	7,303	0.17%	0.66%	3.81	6
Aalborg	213,558	143km	382	10,994	0.29%	0.98%	3.39	7
Horsens	89,598	82km	203	6,872	0.18%	0.62%	3.39	8
Aarhus	340,421	106km	5,419	81,090	3.33%	7.20%	2.16	23
Ringkøbing-Skjern	57,005	0km	90,355	411,494	54.52%	36.50%	0.67	68



(a) From Ringkøbing-Skjern.

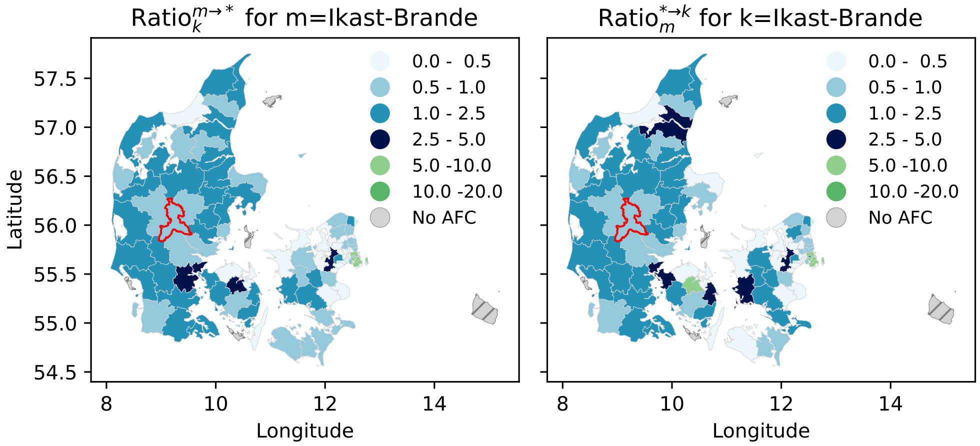
Figure B.9: Ratio between the conditional probability of travelling and searching in relation of Ringkøbing-Skjern.

Table B.22: Top Ratio$_k^{m \to *}$ with the four largest municipalities and Ikast-Brande - Condition from the municipality Ikast-Brande.

To	Population	Distance	Trips	Searches	$\widetilde{pi}_k^{m->*}$	$\widetilde{\lambda}_k^{m->*}$	Ratio$_k^{m->*}$	Rank
Tårnby	43,063	218km	30	8,441	0.11%	0.96%	8.43	1
København	613,288	211km	841	50,437	0.83%	5.68%	6.89	2
Odense	202,348	102km	504	21,663	0.53%	2.45%	4.62	3
Roskilde	87,382	185km	82	4,269	0.16%	0.49%	3.08	4
Fredericia	51,326	59km	416	10,700	0.45%	1.21%	2.69	5
Kolding	92,515	66km	442	10,877	0.48%	1.23%	2.60	6
Middelfart	38,210	77km	83	3,457	0.16%	0.40%	2.49	7
Aalborg	213,558	118km	468	9,240	0.50%	1.05%	2.11	12
Aarhus	340,421	59km	8,611	98,230	7.64%	11.06%	1.45	25
Ikast-Brande	41,191	0km	18,067	112,749	15.93%	12.69%	0.80	45

Table B.23: Top Ratio$_m^{* \to k}$ with the four largest municipalities and Ikast-Brande - Condition to municipality Ikast-Brande.

To	Population	Distance	Trips	Searches	$\widetilde{\pi}_k^{*->k}$	$\widetilde{\lambda}_k^{*->k}$	Ratio$_k^{*->k}$	Rank
København	613,288	211km	831	50,843	0.81%	6.09%	7.48	1
Tårnby	43,063	218km	56	7,792	0.14%	0.94%	6.91	2
Odense	202,348	102km	504	23,616	0.53%	2.84%	5.37	3
Slagelse	78,968	153km	95	5,603	0.17%	0.68%	4.00	4
Fredericia	51,326	59km	406	11,746	0.44%	1.42%	3.20	5
Roskilde	87,382	185km	90	3,846	0.17%	0.47%	2.84	6
Nyborg	32,032	124km	42	2,773	0.12%	0.34%	2.76	7
Aalborg	213,558	118km	368	9,241	0.41%	1.12%	2.73	8
Aarhus	340,421	59km	8,949	99,502	7.92%	11.91%	1.50	25
Ikast-Brande	41,191	0km	18,067	112,749	15.89%	13.49%	0.85	48



(a) From Ikast-Brande.

Figure B.10: Ratio between the conditional probability of travelling and searching in relation of Ikast-Brande.

Bibliography

Azalden Alsger, Behrang Assemi, Mahmoud Mesbah, and Luis Ferreira. Validating and improving public transport origin–destination estimation algorithm using smart card fare data. *Transportation Research Part C: Emerging Technologies*, 68:490–506, 7 2016. ISSN 0968090X. doi: 10.1016/j.trc.2016.05.004. URL `https://linkinghub.elsevier.com/retrieve/pii/S0968090X16300353`.

Azalden Alsger, Ahmad Tavassoli, Mahmoud Mesbah, Luis Ferreira, and Mark Hickman. Public transport trip purpose inference using smart card fare data. *Transportation Research Part C: Emerging Technologies*, 87:123–137, 2 2018. ISSN 0968090X. doi: 10.1016/j.trc.2017.12.016. URL `https://linkinghub.elsevier.com/retrieve/pii/S0968090X17303777`.

James J. Barry, Robert Freimer, and Howard Slavin. Use of Entry-Only Automatic Fare Collection Data to Estimate Linked Transit Trips in New York City. *Transportation Research Record: Journal of the Transportation Research Board*, 2112(1):53–61, 1 2009. ISSN 0361-1981. doi: 10.3141/2112-07. URL `http://journals.sagepub.com/doi/10.3141/2112-07`.

Thomas Bayes. LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr.

Price, in a letter to John Canton, A. M. F. R. S. *Philosophical Transactions of the Royal Society of London*, 53:370–418, 12 1763. ISSN 0261-0523. doi: 10.1098/rstl.1763.0053. URL https://royalsocietypublishing.org/doi/10.1098/rstl.1763.0053.

José M Bernardo. Modern Bayesian Inference: Foundations and Objective Methods. In Prasanta S Bandyopadhyay and Malcolm R Forster, editors, *Philosophy of Statistics*, volume 7 of *Handbook of the Philosophy of Science*, pages 263–306. North-Holland, Amsterdam, 2011. doi: https://doi.org/10.1016/B978-0-444-51862-0.50008-3. URL https://www.sciencedirect.com/science/article/pii/B9780444518620500083.

Michael Betancourt. A Conceptual Introduction to Hamiltonian Monte Carlo. 1 2017. URL http://arxiv.org/abs/1701.02434.

Christopher M Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006. ISBN 0387310738.

Fabrizio Cerreto, Bo Friis Nielsen, Otto Anker Nielsen, and Steven S. Harrod. Application of Data Clustering to Railway Delay Pattern Recognition. *Journal of Advanced Transportation*, 2018:1–18, 2018. ISSN 0197-6729. doi: 10.1155/2018/6164534. URL https://www.hindawi.com/journals/jat/2018/6164534/.

W. Daamen, P. H.L. Bovy, and S. P. Hoogendoorn. Choices between stairs, escalators and ramps in stations. In *10th International Conference on Computer System Design and Operation in the Railway and Other Transit Systems, COMPRAIL 2006, CR06*, pages 3–12, 2006. ISBN 1845641779. doi: 10.2495/CR060011.

Danmarks Statisk. Statbank.dk - Tabel FOLK1A and RAS301, 2021. URL https://www.statbank.dk.

Malvika Dixit, Ties Brands, Niels van Oort, Oded Cats, and Serge Hoogendoorn. Passenger Travel Time Reliability for Multimodal

Public Transport Journeys. *Transportation Research Record*, 2673(2): 149–160, 2019. ISSN 21694052. doi: 10.1177/0361198118825459.

DOT. Zonekort Over Sjælland - Din Offentlige Trafik, 2021. URL `https://dinoffentligetransport.dk/trafikinfo/trafikkort/zonekort/`.

Mohamed K. El Mahrsi, Etienne Come, Latifa Oukhellou, and Michel Verleysen. Clustering Smart Card Data for Urban Mobility Analysis. *IEEE Transactions on Intelligent Transportation Systems*, 2017. ISSN 15249050. doi: 10.1109/TITS.2016.2600515.

Andrew Gelman and Christian Hennig. Beyond subjective and objective in statistics. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180(4):967–1033, 10 2017. ISSN 09641998. doi: 10.1111/rssa.12276. URL `https://onlinelibrary.wiley.com/doi/10.1111/rssa.12276`.

Andrew Gelman and Deborah Nolan. You Can Load a Die, But You Can't Bias a Coin. *The American Statistician*, 56(4):308–311, 2002. doi: 10.1198/000313002605. URL `https://doi.org/10.1198/000313002605`.

Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 11 2013. ISBN 9780429113079. doi: 10.1201/b16018. URL `https://www.taylorfrancis.com/books/9781439898208`.

Andrew Gelman, Daniel Simpson, and Michael Betancourt. The Prior Can Often Only Be Understood in the Context of the Likelihood. *Entropy*, 19(10):555, 10 2017. ISSN 1099-4300. doi: 10.3390/e19100555. URL `http://www.mdpi.com/1099-4300/19/10/555`.

Zoubin Ghahramani. Unsupervised Learning. In *Advanced Lectures on Machine Learning*, pages 72–112. Springer Berlin Heidelberg, 2004. ISBN 978-3-540-28650-9. doi: 10.1007/978-3-540-28650-9{_}5. URL `http://link.springer.com/10.1007/978-3-540-28650-95`.

Zoubin Ghahramani. Bayesian non-parametrics and the probabilistic approach to modelling. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1984):20110553, 2 2013. ISSN 1364-503X. doi: 10.1098/rsta.2011.0553. URL `https://royalsocietypublishing.org/doi/10.1098/rsta.2011.0553`.

Zoubin Ghahramani. Probabilistic machine learning and artificial intelligence. *Nature*, 521(7553):452–459, 5 2015. ISSN 0028-0836. doi: 10.1038/nature14541. URL `http://www.nature.com/articles/nature14541`.

W Keith Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

Li He and Martin Trépanier. Estimating the destination of unlinked trips in transit smart card fare data. *Transportation Research Record*, 2535:97–104, 2015. ISSN 03611981. doi: 10.3141/2535-11.

Matthew D. Hoffman and Andrew Gelman. The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 2014. ISSN 15337928.

Sung-Pil Hong, Yun-Hong Min, Myoung-Ju Park, Kyung Min Kim, and Suk Mun Oh. Precise estimation of connections of metro passengers from Smart Card data. *Transportation*, 43(5):749–769, 9 2016. ISSN 0049-4488. doi: 10.1007/s11116-015-9617-y. URL `http://link.springer.com/10.1007/s11116-015-9617-y`.

Christina Iliopoulou and Konstantinos Kepaptsoglou. Combining ITS and optimization in public transportation planning: state of the art and future research paths. *European Transport Research Review*, 11(1): 27, 12 2019. ISSN 1867-0717. doi: 10.1186/s12544-019-0365-5. URL `https://link.springer.com/article/10.1186/s12544-019-0365-5`.

Jesper Bláfoss Ingvardson, Otto Anker Nielsen, Sebastián Raveau, and Bo Friis Nielsen. Passenger arrival and waiting time distributions dependent on train service frequency and station characteristics: A smart card data analysis. *Transportation Research Part C: Emerging Technologies*, 90(September 2017):292–306, 5 2018. ISSN 0968090X. doi: 10.1016/j.trc.2018.03.006. URL `https://linkinghub.elsevier.com/retrieve/pii/S0968090X1830319X`.

Jian Gang Jin, Loon Ching Tang, Lijun Sun, and Der-Horng Lee. Enhancing metro network resilience via localized integration with bus services. *Transportation Research Part E: Logistics and Transportation Review*, 63:17–30, 3 2014. ISSN 13665545. doi: 10.1016/j.tre.2014.01.002.

Pramesh Kumar, Alireza Khani, and Qing He. A robust method for estimating transit passenger trajectories using automated data. *Transportation Research Part C: Emerging Technologies*, 95(August): 731–747, 10 2018. ISSN 0968090X. doi: 10.1016/j.trc.2018.08.006. URL `https://doi.org/10.1016/j.trc.2018.08.006https://linkinghub.elsevier.com/retrieve/pii/S0968090X18301633`.

Fumitaka Kurauchi and Jan-Dirk Schmöcker. *Public Transport Planning with Smart Card Data*. CRC Press/Taylor & Francis Group, 1st edition, 2017. ISBN 978-1-4987-2658-0.

Catherine T Lawson, Paul Tomchik, Alex Muro, and Eric Krans. Translation software: An alternative to transit data standards. *Transportation Research Interdisciplinary Perspectives*, 2:100028, 9 2019. ISSN 25901982. doi: 10.1016/j.trip.2019.100028. URL `https://doi.org/10.1016/j.trip.2019.100028https://linkinghub.elsevier.com/retrieve/pii/S2590198219300284`.

Tian Li, Dazhi Sun, Peng Jing, and Kaixi Yang. Smart Card Data Mining of Public Transport Destination: A Literature Review. *Information*, 9 (1):18, 1 2018. ISSN 2078-2489. doi: 10.3390/info9010018. URL `http://www.mdpi.com/2078-2489/9/1/18`.

Kai Lu, Alireza Khani, and Baoming Han. A trip purpose-based data-driven alighting station choice model using transit smart card data. *Complexity*, 2018, 2018. ISSN 10990526. doi: 10.1155/2018/3412070.

Ding Luo, Loïc Bonnetain, Oded Cats, and Hans van Lint. Constructing Spatiotemporal Load Profiles of Transit Vehicles with Multiple Data Sources. *Transportation Research Record: Journal of the Transportation Research Board*, 2672(8):175–186, 12 2018. ISSN 0361-1981. doi: 10.1177/0361198118781166. URL `http://journals.sagepub.com/doi/10.1177/0361198118781166`.

Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, 21(6): 1087–1092, 1953. doi: 10.1063/1.1699114. URL `https://doi.org/10.1063/1.1699114`.

Yun-Hong Min, Suk-Joon Ko, Kyung Min Kim, and Sung-Pil Hong. Mining missing train logs from Smart Card data. *Transportation Research Part C: Emerging Technologies*, 63:170–181, 2 2016. ISSN 0968090X. doi: 10.1016/j.trc.2015.11.015. URL `https://linkinghub.elsevier.com/retrieve/pii/S0968090X15004155`.

Neema Nassir, Alireza Khani, Sang Gu Lee, Hyunsoo Noh, and Mark Hickman. Transit Stop-Level Origin–Destination Estimation through Use of Transit Schedule and Automated Data Collection System. *Transportation Research Record: Journal of the Transportation Research Board*, 2263(1):140–150, 1 2011. ISSN 0361-1981. doi: 10.3141/2263-16. URL `http://journals.sagepub.com/doi/10.3141/2263-16`.

Radford M Neal. An Improved Acceptance Procedure for the Hybrid Monte Carlo Algorithm. *Journal of Computational Physics*, 111(1): 194–203, 1994. ISSN 0021-9991. doi: https://doi.org/10.1006/jcph.1994.1054. URL `https://www.sciencedirect.com/science/article/pii/S0021999184710540`.

Radford M. Neal. MCMC using Hamiltonian dynamics. 2012. doi: arXiv:1206.1901.

Panagiotis Papastamoulis. label.switching : An R Package for Dealing with the Label Switching Problem in MCMC Outputs. *Journal of Statistical Software*, 69(Code Snippet 1), 2016. ISSN 1548-7660. doi: 10.18637/jss.v069.c01. URL `http://www.jstatsoft.org/v69/c01/`.

Jens Parbo, Otto Anker Nielsen, and Carlo Giacomo Prato. User perspectives in public transport timetable optimisation. *Transportation Research Part C: Emerging Technologies*, 48:269–284, 2014. ISSN 0968090X. doi: 10.1016/j.trc.2014.09.005.

Marie-Pier Pelletier, Martin Trépanier, and Catherine Morency. Smart card data use in public transit: A literature review. *Transportation Research Part C: Emerging Technologies*, 19(4):557–568, 8 2011. ISSN 0968090X. doi: 10.1016/j.trc.2010.12.003. URL `https://linkinghub.elsevier.com/retrieve/pii/S0968090X1000166X`.

Rejsekort & Rejseplan A/S. Rejsekort i tal, 2021a. URL `https://www.rejsekort.dk/da/rkrp/rejsekort--rejseplan-i-tal`.

Rejsekort & Rejseplan A/S. Zonekort Danmark, 2021b. URL `https://www.rejsekort.dk/da/hjaelp/priser`.

Steve Robinson, Baskaran Narayanan, Nelson Toh, and Francisco Pereira. Methods for pre-processing smartcard data to improve data quality. *Transportation Research Part C: Emerging Technologies*, 49: 43–58, 2014. ISSN 0968090X. doi: 10.1016/j.trc.2014.10.006. URL `http://dx.doi.org/10.1016/j.trc.2014.10.006`.

J.J. van Roosmalen. *Forecasting bus ridership with trip planner usage data : a machine learning application*. PhD thesis, University of Twente, 2019. URL `http://data.openov.nl/docs/Roosmalen_MA_BMS.pdf`.

Gabriel E. Sánchez-Martínez. Inference of Public Transportation Trip Destinations by Using Fare Transaction and Vehicle Location Data.

Transportation Research Record: Journal of the Transportation Research Board, 2652(1):1–7, 1 2017. ISSN 0361-1981. doi: 10.3141/2652-01. URL `http://journals.sagepub.com/doi/10.3141/2652-01`.

Matthew Stephens. Dealing with Label Switching in Mixture Models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 62(4):795–809, 2000. ISSN 13697412, 14679868. URL `http://www.jstor.org/stable/2680622`.

Hong En Tan, De Wen Soh, Yong Sheng Soh, and Muhamad Azfar Ramli. Derivation of train arrival timings through correlations from individual passenger farecard data. *Transportation*, 1 2021. ISSN 0049-4488. doi: 10.1007/s11116-021-10164-w. URL `http://link.springer.com/10.1007/s11116-021-10164-w`.

Niels van Oort, Ties Brands, and Erik de Romph. Short-Term Prediction of Ridership on Public Transport with Smart Card Data. *Transportation Research Record: Journal of the Transportation Research Board*, 2535(1):105–111, 1 2015. ISSN 0361-1981. doi: 10.3141/2535-12. URL `http://journals.sagepub.com/doi/10.3141/2535-12`.

Ziyulong Wang. *Predicting Short-term Bus Ridership with Trip Planner Data: A Machine Learning Approach*. PhD thesis, TU Delft, 2020. URL `https://repository.tudelft.nl/islandora/object/uuid:f1e4b495-d2ad-4a1e-803e-13e6c9b39f4a`.

Timothy F. Welch and Alyas Widita. Big data in public transportation: a review of sources and methods. *Transport Reviews*, 0(0):1–24, 2019. ISSN 14645327. doi: 10.1080/01441647.2019.1616849. URL `https://doi.org/10.1080/01441647.2019.1616849`.

M.D. Yap, O. Cats, N. van Oort, and S.P. Hoogendoorn. A robust transfer inference algorithm for public transport journeys during disruptions. *Transportation Research Procedia*, 27:1042–1049, 2017. ISSN 23521465. doi: 10.1016/j.trpro.2017.12.099. URL `https://linkinghub.elsevier.com/retrieve/pii/S2352146517309961`.

Jinhua Zhao, Michael Frumin, Nigel Wilson, and Zhan Zhao. Unified estimator for excess journey time under heterogeneous passenger incidence behavior using smartcard data. *Transportation Research Part C: Emerging Technologies*, 34:70–88, 9 2013. ISSN 0968090X. doi: 10.1016/j.trc.2013.05.009. URL `https://linkinghub.elsevier.com/retrieve/pii/S0968090X13001101`.