UNIVERSITY OF TARTU
Faculty of Science and Technology
Institute of Computer Science
Software Engineering Curriculum

Erald Keshi

Alighting Estimation in Entry-Only AFC Systems; a Case Study of Tartu City

Master's Thesis (30 ECTS)

Supervisors: Mozhgan Pourmoradnasseri, Ph.D.
Amnir Hadachi, Ph.D.

Tartu 2023

Alighting Estimation in Entry-Only AFC Systems; a Case Study of Tartu City

Abstract:

For planning an efficient and ecological public transport network accurate understanding of citizens' mobility patterns is essential. A deep understanding of these mobility patterns can only be achieved through the analysis of both boarding and alighting trip information. Automatic fare collection (AFC) systems have become a popular data source for public transportation research, but entry-only AFC systems, including the one used in Tartu, do not capture critical data such as alighting stations or alighting times. This limits the creation of origin-destination matrices and gaining insights such as bus occupancy rate and peak hours. Probabilistic estimation methods are one approach that could be used to tackle this limitation. The contribution of this thesis is the exploration, development and comparison of different methods for estimating alighting information from entry-only AFC systems. These methods can be used to fill the gaps in data and provide more comprehensive insights into the usage patterns of public transportation, informing better decision-making processes related to route and schedule planning.

Keywords:
Destination estimation model, OD matrix, Automatic fare collection, mobility modeling, Probabilistic estimation methods.

CERCS: P170 (Computer science, numerical analysis, systems, control)

Väljumise hindamine ainult sisenemisega AFC-süsteemides; Tartu juhtumiuuring

Lühikokkuvõte:

Tõhusa ja ökoloogilise ühistranspordivõrgu kavandamiseks on oluline kodanike liikumismustrite täpne mõistmine. Nende ühistranspordi liikumisharjumuste põhjalik mõistmine on võimalik ainult sisenemise ja väljumise andmete analüüsi abil. Automaatsed piletihindade kogumise süsteemid on muutunud populaarseks andmeallikaks ühistranspordi uurimisel, kuid ainult sisenemisega seotud automaatsed piletihindade kogumise süsteemid, nagu näiteks Tartus kasutatav süsteem, ei hõlma selliseid kriitilisi andmeid nagu reisijate sihtkohad või ühistranspordist väljumiste ajad. See piirab lähte- ja sihtkohtade maatriksite koostamist ja selliste andmete saamist nagu busside täituvus ja tipptunnid. Üks võimalus selle piirangu kõrvaldamiseks on kasutada tõenäosuslikke hindamismeetodeid. Käesoleva lõputöö eesmärk on uurida, arendada ja võrrelda erinevaid meetodeid, mille abil saab hinnata ainult sisenemisega seotud AFC-süsteemidest pärit väljumise andmeid. Neid meetodeid saab kasutada andmete puudujääkide täitmiseks ja ulatuslikuma ülevaate andmiseks ühistranspordi kasutusmustritest, mis aitab paremini otsustada marsruutide ja sõiduplaanide planeerimise üle.

Võtmesõnad:

Sihtkoha hindamise mudel, Algpunkt-sihtpunkt Maatriks, Automaatne piletihindade kogumine, liikuvuse modelleerimine, Tõenäosusliku hindamise meetodid.

CERCS: P170: Arvutiteadus, arvutusmeetodid, süsteemid, juhtimine (automaatjuhtimisteooria)

Contents

1 Introduction

Urban transportation mobility is a critical aspect of modern and sustainable cities. It plays a significant role in shaping the quality of life of citizens by providing them with access to different services, employment, and social activities. Therefore, understanding where and how citizens use the public transportation system while performing their daily activities could lead to a more efficient and integrated bus network. To this end, researchers have been utilizing a variety of methods, ranging from user surveys to making use of digital data while trying to model behavioral patterns for different cities. The results could be used for better route and schedule planning or bus transport reform validation.

While user surveys have been a traditional method for collecting data on public transportation usage, they often take several years to complete, and the data collected may not be comprehensive or representative of the entire population. As a result, researchers have been exploring the use of different data sources, such as automatic fare collection (AFC) systems, to gain insights into the usage patterns of public transportation.

The aim of this master's thesis is to explore the different data available for the city of Tartu and use it to gain insights into how the bus network is being used by the citizens. By leveraging the available data, it is possible to develop a better understanding of the mobility patterns of citizens, which could inform decision-making processes related to route and schedule planning, and ultimately contribute to a more efficient and reliable bus network.

1.1 Problem statement

The use of AFC systems is becoming increasingly popular in public transportation, with most cities adopting entry-only AFC systems. The city of Tartu is no exception, as it relies solely on entry-only AFC systems where passengers validate their tickets while boarding the bus. However, these type of systems do not capture critical information such as alighting station or alighting time, which limits the creation of origin-destination matrices at the stop or district level and gaining insights such as bus occupancy rate at any point in time. An origin-destination (OD) matrix is a data representation used in transportation analysis, showing the flow of passengers or vehicles between various origins and destinations within a specified area.

Furthermore, not all AFC entry systems collect the same data, and the agencies that own this data can be reluctant to share sensitive information with researchers, such as user identifiers or user bus card identifiers. As a result, there is a need to explore alternative data sources or develop methods to fill the gaps in data, which could potentially provide more comprehensive insights into the usage patterns of public transportation and inform better decision-making processes related to route and schedule planning.

Therefore, the main problem addressed in this master's thesis is how to leverage the available data and develop methods to overcome the limitations of entry-only AFC systems in the city of Tartu.

1.2 Contribution

The main contribution of this master's thesis is the exploration, development and implementation of methods for estimating alighting information from entry-only AFC systems, using the data collected from the AFC system installed in Tartu city buses. The general limitations of entry-only AFC systems, which do not capture important data such as alighting station or alighting time, are tackled by implementing several methods for estimating this missing information. The validation of the accuracy of the implemented methods is a challenge in itself with the lack of abundant real-world alighting infromation. To address this challenge, two custom validation methods based on counter data and boarding data are developed, which enables the comparison of different methods. Further one of the top performing methods is selected to estimate alighting information for the available dataset and is used to construct the origin-destination (OD) matrix at the district level. The results are presented in a public web-based dashboard.

1.3 Roadmap

The rest of this thesis is organized as follows:

Chapter 2 (Background): This chapter describes the research done for selecting the alighting information for different public transportation systems. It explains how the AFC systems used can impact the solution and why the solution is highly influenced by the available data.

Chapter 3 (Data): This chapter describes the AFC data pertaining to the city of Tartu and showcases the outcomes of the exploratory data analysis, which served as a guide in formulating the presented solution.

Chapter 4 (Methodology): This chapter begins by describing the assumptions which underpin the proposed solution. Further it explains in detail the probability-based methods used for predicting alighting information and discusses the limitations of this methodology and the reasons it was selected for this work. Next, the process for validating the probability-based methods is presented. Lastly, the inner workings of the pipeline is explained.

Chapter 5 (Experiments): This chapter begins with providing some experiment-based evidence that support the assumptions made in the methodology chapter. Subsequently, the results obtained from both synthetic and real-world validation of the four distinct probability-based algorithms utilized for alighting station estimation are presented. A comprehensive comparison among these methods is conducted, followed by a thorough interpretation of the findings.

Chapter 5: (Discussions) In this chapter, the outcomes of the experiments and the main derived lessons are discussed. Based on these discussions, a roadmap for future work is provided.

2 Background

This chapter will provide the necessary information for understanding the scope and the terminologies of this project.

The first section will describe what AFC systems are, how they work and which data they typically collect. The second section describes the bus transportation system of Tartu and the AFC system it uses. The third section gives an overview of the current research regarding entry-only AFC data.

2.1 Tracking boarding and alighting in public transport

For better planning and understanding of bus system passenger mobility, boarding and alighting information are crucial. Throughout the years, several different methods have been used to obtain this information. Some of these methods include labour-intensive checks, where a human operator would manually count passengers getting on and off at each stop. More modern methods include the use of Automatic Passenger Counters (APC). These are electronic devices installed on public transport vehicles such as buses, trains, and trams to accurately count the number of passengers boarding and alighting at each stop.These devices use various technologies such as infrared sensors, video cameras, or weight sensors to count the number of passengers who enter or exit the vehicle. This data is then transmitted wirelessly to a central database, where it can be analyzed to gain insights into passenger demand, service performance, and revenue management. However, APCs are still far from widespread use [12]. Due to their cost, they typically get installed at a sample of the buses, or only for a limited period of time as it is pretty expensive to install them in the entire bus fleet.

AFC systems are another way that can be used to retrieve passenger location data. Although their primary goal is automating and simplifying the fare processes for passengers and operators, due to the data they collect AFC systems have been heavily used by researchers to obtain information about boarding and alighting behaviour. Generally, AFC systems are able to produce transaction level information for each passenger validation, including the time validation took place and also the location, stop or trip information.

Based on the implementation of AFC systems they can be split into two main groups:

1. **Entry-exit AFC systems**. These systems require passengers to validate their entry and exit. This functionality is commonly employed in subway settings, resembling the turnstile mechanisms that necessitate scanning upon entering or leaving a station.

2. **Entry-only AFC systems**. Generally due to passenger experience and ease of use most bus transportation systems in the world are using entry-only automatic

fare collection systems. These types of systems only require passengers to validate while boarding the bus, meaning that they are unable to collect additional information about alighting stations.

2.2 Tartu's bus transportation system

Tartu, a city in Estonia with a population of 97,435 relies on buses as its only method of public transportation. In September 2015, Tartu implemented a modern AFC system in its bus network [2], which enables passengers to use a contactless plastic chip card (Fig 1) or a similar-looking sticker to pay for their fares.

When passengers board the bus, they must validate their card by tapping it on one of the validators (Fig 2) located inside the bus. The fare is automatically deducted from the card balance at the moment of validation. The ticket is valid for a one-hour journey, and if a passenger validates their ticket again within the one-hour window, no further funds will be deducted from the card balance. The AFC system also allows debit or credit cards with contactless capabilities, as well as student cards, to be used as payment methods. This system also requires that passengers with the right to travel for free should use their cards to validate their journeys.

Failure to validate the ticket when boarding the bus, even if the card has an active purchased ticket, may result in a fine for the passenger.

The city also offers more payment opportunities besides contactless payments. One example is the purchase of an 1 hour QR ticket sent to passenger's email [3]. In such cases it is not possible to validate these tickets and they will not be part of the AFC system's data.

Figure 1. Tartu bus card [2].

Figure 2. Tartu bus validator [2].

2.3 A literature review for entry-only AFC destination estimation

Many researchers have tried to tackle the problem of missing destination information on entry only AFC systems. Several methods have been used depending on the data available and main goal of the research.These methods can be grouped in 3 main groups [11] :

2.3.1 Trip chaining models

Trip chaining model is a method used to recreate the route that passengers have followed throughout the day [5], [16], [4]. If a passenger boarded the first bus at Stop A and then boarded a second bus at Stop C, it can be said under specific assumptions that the passenger exited the first bus at a stop near Stop C. To make this model work, it is crucial for the AFC system data to include the identifier of each passenger's card. This allows multiple trips of the same passenger to be chained together within a day.

As briefly mentioned the trip chaining model is used under several important assumptions:

1. There is no other mode of transportation that the passenger uses between his trips [16].
2. Travellers will not walk a long distance from one stop to another [16].
3. Passengers' last trip of the day ends at the same station where its first trip started [5].

2.3.2 Probability based models

Probability based methods use several data such as travel distance, passenger number, land use or transfer capacity to estimate the alighting station [11]. Several examples of these methodologies include [6], [17], [15] and [14]. These types of methods are good for estimating high level picture of mobility, but not accurate enough on single validation level.

2.3.3 Deep learning model

Destination estimation through the use of deep learning models has also been explored from researchers. For these methods it is important to mention that although the models itself work for entry-only AFC systems, for the training dataset it is necessary to also have alighting information. Yu Jie applied a modified BP artificial neural network to estimate the bus OD matrix in China. The model used the boarding number as input and alighting number as output, and was trained using six groups of investigated bus OD data [8]. Meanwhile, Jaeyoung Jung et al. [9] estimated the alighting number using smart-card and land data, with the help of a deep learning model.

2.3.4 Comparison between different models

As summarized in [11], the trip chaining model has the advantage of requiring only smart card data and has a relatively simple algorithm compared to the other 2 methods. It can also forecast the alighting of each passenger, however the algorithm requires having access to a passenger identifier across trips, which is not always possible to obtain due to privacy concerns. In addition the model validation is usually a difficult process. The probability models take into account more comprehensive factors, but it can only estimate the total on-off number of passengers. The deep learning models have very comprehensive considerations and can infer the alight station on an individual passenger basis. However, it requires access to alighting information for training the model and requires abundant data, which can be difficult to obtain [11].

3 Data

First, the data used in this work in presented and explained. Next, an exploratory data analysis (EDA) is performed with the goal of detecting the type and quality of the data, as well as any limitations or biases that may affect the study. The EDA is focused on the most important aspects of the data which affected the proposed solutions and validation methods.

3.1 Available data

Most of the input data comes directly from the company which provides AFC capabilities for Tartu's bus transportation system. Through the use of a public Authenticated API based on HTTP protocol, five different endpoints are used to retrieve JSON data for a specific day provided as parameter. In the following sections the different models present in the dataset are described.

3.1.1 Ticket Validations

The ticket validation data contains information about each ticket validation event that happened throughout the day on each bus. Inspecting a row in the validation file, the most important information present are: timestamp of the validation, the number of the bus line, the ID of the trip, the passengers count and the stop sequence. The stop sequence numbers begin with 1 for the first stop of the trip and increment by 1 for each stop until the last stop of the trip. The validation row also contains fields about card identifiers and document number. However, these fields are empty due to the sensitivity of the values they contain. Below is a sample of a validation row:

```
Timestamp: 2022-01-04 04:31:42
Line: 21
Trip ID: 847273
Passenger count: 1
Stop sequence: 2
Location: 437TNS
Product ID: 6276
Card ID: NaN
Document number: NaN
```

3.1.2 Trips

Each row of the trips data represents an actual trip by a bus which has followed a specific route in a specific direction for a start and end time. The same route and direction can have multiple trips throughout the day. So for a route A, there may exist three different directions: A->B, B->A and B->C->A. In this case, as it will become important later the A->B and both B->A and B->C->A are considered opposite directions. Below is a sample of a trip row:

```
Trip ID: 6359441
Route ID: 369737
Trip short name: Ööliin (P)
Direction ID: A > C
Departure time: 2023-04-30 05:30:00
Arrival time: 2023-04-30 06:35:00
Tour number: 1172329
```

3.1.3 Route

The route data contains all the bus routes in the city for each day. It is important to note that a route does not contain directional information, instead directional information is included in the trip level as mentioned above. Below is a sample of a route row:

```
Route ID: 369737
Route long name: Ööliin (P)
Route short name (Line number): 9
```

3.1.4 Stops

Stops data, contains all the bus stop stations of the city. Each row contains a unique identifier of a stop, its code and the stop name.Below is a sample of a stop row:

```
Stop ID: 4345145
Stop code: 7820249-1
Stop name: Raeplats
```

3.1.5 Punctuality

The purpose of the punctuality data is to give very detailed information about how punctual the bus was in reaching each stop in its planned time. For every trip and every stop in the trip, it contains the time the bus reached that stop. Besides the time information it also shows whether the bus had to stop in each stop; either for passengers to come in or to go out. This piece of information proved to be very useful for the upcoming modeling pipeline. Occasionally the punctuality row may also have values for **counter in** and **counter out** fields. The frequency of this information is explored in section 3.2.2. This information will be of important use while performing real world evaluation. Below is a sample of a punctuality row:

```
Route short name: 21
Trip direction: A>B
Stop sequence: 37
Stop name: Annemõisa
Stop code: 7820017-1
Planned stop arrival: 2022-01-04 05:20:00
Actual arrival: 2022-01-04 05:19:50
Bus stopped lat: 58.372286
Bus stopped lon: 26.780043
Trip ID: 6359461
Counter in: 20
Counter out: 30
Validations count: 16
```

3.1.6 Tartu districts shape data

This static data contains the geographic shape and location of Tartu city districts. Below is a sample row for Annelinna district:

```
Object ID: 21
District name: Annelinna
Population: 24551
Shape.STAr: 5474557.366
Shape.STLe: 10626.34457
Geometry: POLYGON ((26.69126249187359
... 58.40717651396527))
```

The district shape data is necessary for the validation. As shown in Section 4.2.3 the mapping between the bus stops to its corresponding district can only be done if the geographical shape and location of the district is known.

3.1.7 Tartu stops from peatus.ee

As obvious from the example in 3.1.4, the information about stops coming from AFC system does not contain the latitude and longitude for each stop. Knowing the latitude and longitude allows to map each station to the district where it is located. The fetching of the location for each stop is achieved with the use of the website [1], which allows the possibility to download a daily General Transit Feed Specification (GTFS) file for the country of Estonia. A row from this dataset contains the below information:

```
Stop Id: 154135
Stop Code: 7820378-1
Stop Name: Kvartsi
Stop Lat: 59.406376
Stop Lon: 28.164083
Stop Area: Tartu linn,
Authority: Tartu LV
```

3.2 Exploratory data analysis

The exploratory data analysis begins with obervations regarding the input data quality. Further a paragraph is dedicated to the APC data.

3.2.1 Data quality issues

In this study, data obtained from the AFC system was utilized. These data have proven to be of high quality. No major gap was noticed in the dataset; however, there were 2 types of discrepancies which are attributed primarily to user behavior rather than technical or data issues.

- The AFC system boarding station is the stop with the highest proximity to the bus at the time of validation. However, there are some instances where the punctuality data for a particular trip indicates that the bus did not stop at a specific station despite ticket validations occurring at that location during the trip. This can be attributed to users forgetting to validate their ticket upon boarding and doing so later in the journey. While such occurrences constitute a small fraction of the data, this issue was addressed by assuming that the user boarded at the first previous stop where the bus had stopped, rather than discarding the data entirely. It is worth noting that the exact boarding station of the passenger is not of utmost importance, as the subsequent methods comparisons and OD matrices are generated on a district level.

- As observed in the exploration of punctuality data, counter in and counter out data are available for all stops for some of the trips during the day, indicating the real number of passengers who boarded or exited the bus at each specific station. The availability of this information highlighted the issue of passengers who do not validate their tickets, as in many cases, the counter in value was found to be significantly higher than the number of validations at that particular stop. This discrepancy was also evident at the trip level, indicating that these are not instances of late validation issue, as discussed earlier, but rather passengers possibly forgetting to validate the tickets during the trip or passengers using another method of payment such as the purchase of a 1-hour QR ticket.

The second data quality issue is not directly addressed in this thesis because both the solution and the validation methods use either the ticket validation information or the counter information in isolation from each other, bypassing the effect of the disrepancy between the two sources towards the end result.

3.2.2 APC data

As noted in the Punctuality section 3.1.5 the dataset contains APC information regarding real counter in and counter out for some trips during most days. Given the importance of this counter information in the real-world validation method a focused data analysis was done to understand the potential weaknesses. Analysis was performed for the data of entire 2022 year and statistical results to better understand these data are presented below.

Table 1. Distribution of APC data for days of 2022 (38 days with 0 data are filtered out)

	Results				
	Mean	**Std deviation**	**Min**	**Max**	**Median**
No. of Unique Trips with APC data	18.45	5.67	1	32	20
Percent of APC trips/number of total trips	1.75%	0.65%	0.08%	4.01%	1.74%
APC counter out	561.73	246.37	11	1361	557
APC counter out > 300	629.48	202.15	300	1361	601

Out of the 365 days of the 2022 year, 38 days had no counter data. The rest of the days had counters information only for a small subset of trips according to the distribution presented in table 1. As visible on the table, after filtering the days which do not have APC data, on average 1.75% of trips are covered with APC data per day. The number of counter out registered as leaving the bus in these trips ranges from 11 to 1361 per day with mean of 561.73 passengers. The last row shows the distribution of the counter out values per day if the days where the counter out value is smaller than 300 are filtered out. The year 2022 contains 49 such days. After these days are filtered out, as expected both mean and median are increased. Upon filtering both 0 counter days and days with less than 300 counts recorded by the APC there are 279 days remaining.

4 Methodology

This section describes the methodology used for alighting station estimation based on the available data.

First, the model group used for alighting station estimation and the assumptions made for this analysis are outlined. Then four different probability algorithms used for alighting station estimation are described, including their strengths and weaknesses.

Finally, the validation methods used to validate the performance of the probability methods are presented. The validation methods will allow to determine the effectiveness of each of the estimation methods in estimating alighting stations on the district level based on the AFC system data, and to compare its performance to other methods.

The last section is dedicated to the pipeline which stands at the core of this thesis by enabling a fully automated processing; starting from the retrieval of input data up until the OD matrices generation.

4.1 Alighting estimation methods

As seen in the data section, due to lack of the passenger identifier data for chaining sequential trips of the same passenger, it was not possible to explore the trip chaining method. Instead, four different probability-based algorithms are explored and compared.

Despite the inability to use trip chaining algorithms directly, the assumptions of these group of models make are used to inform the algorithm development and in the construction of validation methods.

4.1.1 Assumptions

Many studies have found that human mobility patterns are highly regular with individuals frequently returning to the same few locations such as work place and home [13] [7]. These patterns typically follow the daily circadian rhythm in 24 hour cycles. Further [13] by studying the time that individuals in Paris and Chicago spent in specific locations, reported a relatively flat distribution with peaks around 14h of individuals staying at home and two peaks of 3.5h and 8.6h spent at work. The rest of the activities show a lower occurrence with the increase in their duration and are broadly distributed. The studied regularity and understanding of human mobility patterns have been the basis for different assumptions used in alighting station estimation methods. Some of these assumptions are presented below:

1. Passengers' last trip of the day ends at the same station where its first trip started [5].

2. Symmetry assumption; over the course of one day, pattern of daily passenger boardings in one direction mirrors the daily alighting pattern in the opposite direction [12].

3. Midday symmetry assumption refers to the assumption that the pattern of passenger boarding in one direction during the pre-cutoff time is mirrored by the pattern of passenger alighting during the evening. Specifically, if a passenger begins their daily commute in the morning from Stop A, it is assumed that the same passenger will stop at Stop A during their afternoon trip. This assumption was based on [13], where it was implied that people usually have regular round trips starting from the place of residence.

   Later, an investigation is conducted with the aim of providing evidence about the Assumptions 2 and 3 in the city of Tartu.

4.1.2 Alighting candidate stops selection

The process of estimating the alighting stop begins by first creating a shortlist of candidate stops denoted as $candidateStops$. The data exploration step has revealed that the exact stations each bus followed for each trip are known, which eliminates the need to consider temporary route changes if static routes were used. Additionally there is information about the station where passengers have boarded the bus, allowing the exclusion of all stops before (including) the boarding stop. After applying these filters, a list of remaining stops is obtained. To further shorten this list, information from the punctuality file is utilized, which indicates whether a bus has stopped at a particular station during its trip. It was observed that on average 15% of the total visited stops fall under this category, and by removing them, a shorter list of candidate stops is obtained, contributing to more accurate estimations.

As an example Figure 3 illustrates a bus trip and all of its stop stations, where the black dots represent stop stations in the route where bus did not stop while the white dots represent stop stations where bus did stop. Given that a validation occurred when the bus was at stop station number 3, a candidate stops list will be created, where it is possible that the passenger associated with the validation might have alighted.

First, all stops before and including the stop where the validation happened are filtered out. As a result the possible list of stations where the passenger might have alighted is $candidateStops = [4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22]$.

Lastly, all the stop stations where the bus did not stop are also filtered out leaving a shorter candidate stop list of

$$candidateStops = [7, 8, 9, 11, 12, 13, 14, 15, 16, 17, 19, 20, 21, 22].$$

After creating a shortlist of candidate stops each one of the algorithms is applied to

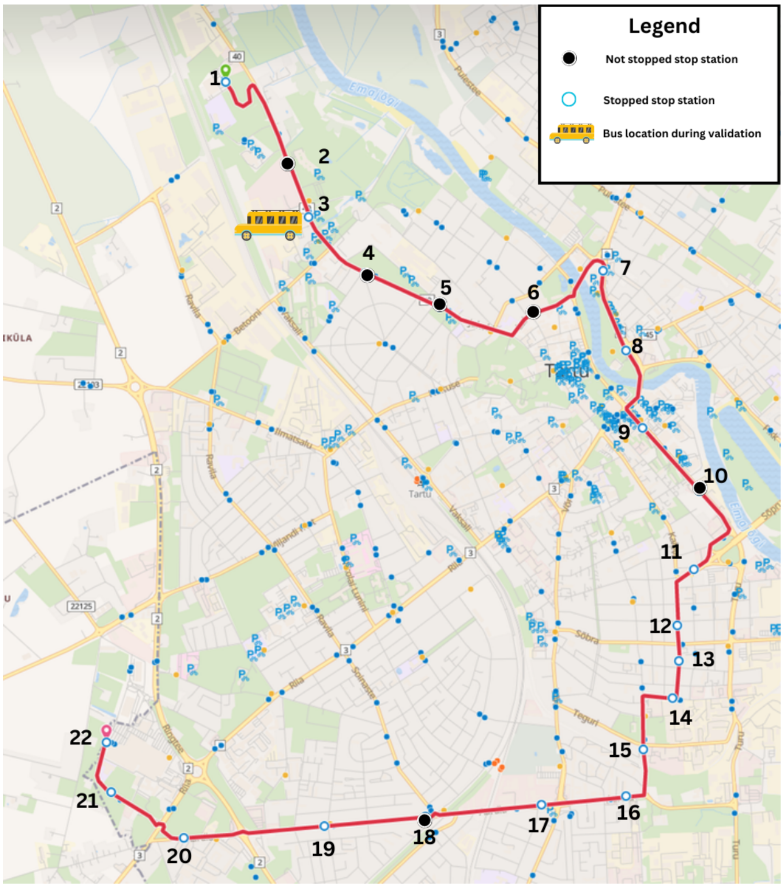


Figure 3. Illustration of candidate stops selection

pick one of the candidates as the destination stop. As shown below the algorithms differ on their assumptions and on their level of complexity.

4.1.3 Uniform distribution

In order to establish a baseline for comparison, the Uniform Distribution probability model is chosen. It assumes that all candidate stops have equal likelihood of being selected as the alighting station. Simply, a random station from the list of candidate stations is selected as the alighting station. This model provides a simple and straightforward approach for comparison with more complex models.

Let n be the number of candidate stops. The probability to pick a stop using uniform distribution is:

$$P(stop_i) = \frac{1}{n} \qquad \text{and} \qquad \sum_{1 \le i \le n} P(stop_i) = 1.$$

4.1.4 Normal distribution

This method is based on the assumption of passengers being less likely to choose the bus for very short trip times, which can be easily done on foot, or for very long ones where the car might be preferred. To reflect this assumption, a normal probability distribution is used to assign probabilities to the candidate stops. Specifically, the probability of stopping at each candidate stop follows a normal distribution curve as illustrated in Figure 4, where the mean is the center of the $candidateStops$ list and standard deviation is 3. This results in candidate stops in the middle having higher probabilities of being destination stops compared to those at the edges.

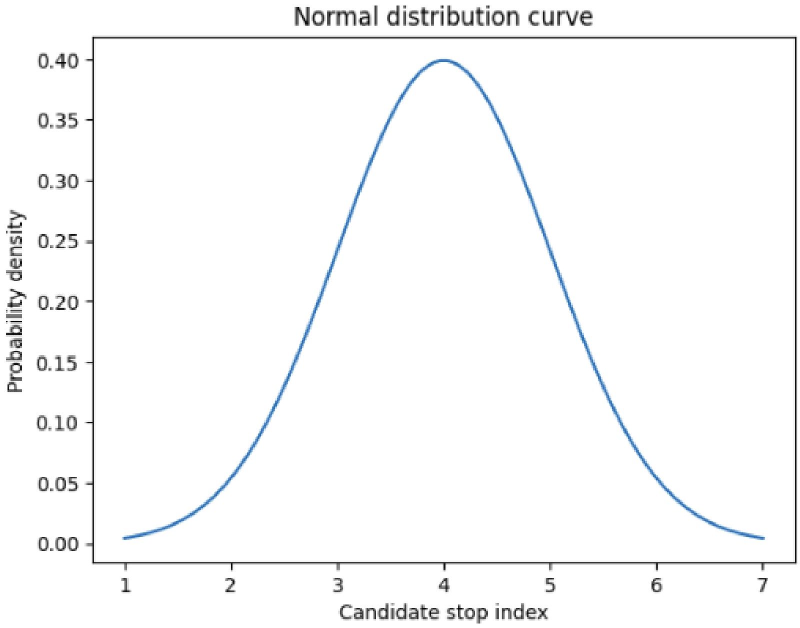


Figure 4. Normal distribution probability curve

4.1.5 Symmetric alighting station estimation

Symmetric alighting station estimation algorithm directly relies on the assumption that starting and ending stops of the day are the same for each passenger (Assumption 1) and indirectly on the symmetry assumption (Assumption 2).

Because passengers are expected to end their day in the same station as they started it can be deduced that the more people boarded at a specific station, the higher the probability that station should be picked as an alighting station throughout the entire day. Also because of symmetry assumption, the alighting patterns and the boarding patterns should match, making the algorithm ensure that across enough validations, the same number of people are alighting at a specific stop station as many people started from it. Given that the origin-destination matrix as well as the validation methods are built on the district level, individual inaccuracies will be flattened, and an accurate result to a reasonable level can be obtained.

As explained in the Algorithm 1 as a first step the number of passengers validations for each stop is counted. Next, in Algorithm 2 each candidate in $candidateStops$ is mapped to the stop's respective popularity count calculated in Algorithm 1. The counts are normalized and the normalized values serve as probabilities of each candidate stop to be picked as an alighting station. Algorithm 2 is executed for each validation in the dataset.

Algorithm 1: Boarding station counts

```
Input: All validations for one day
Result: Count of validations for each station
stopCount ← emptyMap;
for validation in validations do
    validationBoardingStop ← validation[‘stopcode’];
    stopCount[validationBoardingStop]= stopCount[validationBoardingStop]++
return stopCount
```

Algorithm 2: Pick candidate stop based on candidate stop counts

```
Input: All validations for one day
stopCounts ← callAlgorithm1();
for validation in validations do
    candidateStops ← validation[‘candidateStops’];
    candidateStopsRespectiveWeights ← getCounts(candidateStops,
      stopCounts);
    normalizedWeights ← normalizeCounts(candidateStopsRespectiveWeights);
    pickedStop ← choiceWithWeights(candidateStops, normalizedWeights);
    validation[‘alightingStop’] = pickedStop;
```

4.1.6 Midday Symmetric alighting station estimation

In order to further refine the symmetric alighting station estimation method, symmetry assumption within a midday cutoff point (Assumption 3) is indirectly used. Later in section 5.1, some evidence that trips tend to be symmetrical with respect to morning and afternoon periods is provided. This means that if people take the bus in the morning from location X to Y, they are likely to take a bus in the afternoon from Y to X. As such, it is expected that the number of boardings in the morning for one route direction and the number of alightings in the evening for the opposite direction to be similar. The same can be said for the alighting stations in the morning compared to the boarding stations in the evening.

To account for this, a more granular popularity count based on morning and evening validations is created after the validation file is split into two parts: the morning validations and the evening validations. After some experimentation the cutoff time was picked as 14:00:00 local time.

The counting of the popularity of stop stations for each half of the day outputs two different counts maps. The morning counts map tracks all the boarding stops counts for the morning validations, while the evening counts map tracks all the boarding stops counts for the evening validations.

When selecting the exit station from the candidate stops, the counts are cross-referenced; for the afternoon validations, the relative probabilities for each candidate stop is based on the morning counts map, while for the morning validations they are based on the evening counts.

4.2 Validation

To validate the accuracy of the alighting estimation algorithms, it is essential to have ground truth data for boarding and alighting stations. Such data was not abundantly available, and manual collection would have required significant resources. Section 3.2.2 showed that there are APC data for a small subset of trips within the day. When formulating the validation methods, the objective was to leverage a substantial portion of the available data, thus increasing the accuracy of the validation process. For this reason two validation methods were constructed; a synthetic validation method which validates the alighting estimation algorithms based on boarding information only and a counter based validation which uses only the APC real world data.

It is important to note here that although the alighting estimation algorithms operate on the stop level, both validation methods validate them in the district level. There are 2 main reasons this approach was followed.

- As explained in the literature review probabilistic methods are mostly used to gather an high level view of the mobility patterns and they are not expected to be accurate on a fine-grained level.

- OD matrices, which will be the main output of the thesis are typically generated at the district level.

4.2.1 Evaluation metrics

In the thesis several metrics are employed to evaluate the accuracy and performance of the proposed estimation. These metrics include the relative error between two vectors, the weighted relative error between two vectors, the root mean square error (RMSE), and the normalized RMSE.

Vector relative error

The relative error measures the disparity between two vectors by calculating L_2 norm of their difference and dividing it by the L_2 norm of one of the vectors. It is given by the formula:

$$\text{RE} = \frac{\sqrt{\sum_{i=1}^{n}(A_i - B_i)^2}}{\sqrt{\sum_{i=1}^{n} A_i^2}} \tag{1}$$

where:

- n is the number of data points,
- A_i is the actual value at index i,
- B_i is the predicted value at index i.

Weighted vector relative error

The weighted relative error, on the other hand, incorporates varying weights for different elements of the vectors to account for their relative importance. It can be computed using the formula:

$$\text{Weighted Relative Error} = \frac{\sqrt{\sum_{i=1}^{n} w_i \cdot (A_i - B_i)^2}}{\sqrt{\sum_{i=1}^{n} w_i \cdot A_i^2}} \tag{2}$$

where:

- n is the number of data points,
- A_i is the actual value at index i,
- B_i is the predicted value at index i,
- w_i is the weight assigned to data point i.

RMSE

The RMSE measures the average magnitude of the differences between corresponding elements of the vectors and is calculated as follows:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(A_i - B_i)^2} \tag{3}$$

where:

- n is the number of data points,
- A_i is the actual value at index i,
- B_i is the predicted value at index i.

Normalized RMSE

The normalized RMSE provides a standardized measure of the RMSE by dividing it by the range of the target variable. It can be expressed using the formula:

$$\text{NRMSE} = \frac{\text{RMSE}}{\max(A) - \min(A)} \tag{4}$$

where:

- RMSE is the Root Mean Square Error,
- A is the actual data set,
- $\max(A)$ is the maximum value in the data set A,
- $\min(A)$ is the minimum value in the data set A.

4.2.2 Synthetic validation method

This validation method is synthetic because the validation is not performed against ground truth alighting information but rather against assumptions which allow to get a close enough approximation of the ground truth data.

The base premise of this validation method is assumption 1, which states that passengers end their day in the same station where they started it. Logically this assumption also holds true in the district level; passengers end their day in the same district where they started it. So according to this assumption the number of boardings per each district should be exactly the same as the number of alightings for each district. Hence this validation method is focused on finding the relative error between the boardings per district vector and (estimated) alightings per district vector.

The dataset contains complete information regarding all boardings at the station level for a given day. By mapping each station to the corresponding district, and further aggregating them a vector which shows all boardings per district per day is obtained. As per the above deduction, the vector of real alighting information per district per day is expected to be the same as the vector of boardings per district per day. Let's call this vector A_{Real}.

Once one of the probability methods is applied to the candidate stops extracted in section 4.1.2, the alighting stop for each validation in a day is derived. In a manner similar to the real alighting vector, the estimated alighting stop is associated with its corresponding district. After the districts are aggregated, the result is the vector of estimated alighting counts per district per day $A_{Estimated}$.

The synthetic evaluation method per one day E is defined as the relative error between the two vectors A_{Real} and $A_{Estimated}$. For calculating the error the formula 4.2.1 is used.

To evaluate the accuracy of the probability methods over a longer time frame, the evaluation metric E is calculated for multiple days and the average is taken.

Let E_i denote the evaluation metric for day i, where $i = 1, 2, \ldots, k$ is the total number of days. The average evaluation metric can then be calculated for the k days as follows:

$$E_{avg} = (E_1 + E_2 + ... + E_k)/k$$

This gives a measure of the overall accuracy of the algorithm over the entire time period. The standard deviation of the evaluation metric over the k days can also be calculated to get an idea of the variability of the algorithm's performance over time.

4.2.3 Counter based real world validation

To provide real world validation, the counter information present in the punctuality data is utilized. By using the counter out information, the exact number of passengers who exited the bus at each stop can be determined. After estimating the alighting stop for each validation, it is possible to calculate the estimated count out for each stop. The relative error between the estimated count out and the actual count out could be used as the evaluation metric.

However, as outlined in the data quality section 3.2.1, the counter representing the number of passengers boarding the bus at a specific stop (counter in) was significantly larger than the number of validations recorded at the respective stop. Because the estimated count out numbers would be based on the ticket validation numbers, the inherent error from unvalidated tickets would always negatively affect the evaluation metric, despite the estimation efforts.

Because of the discrepancies between the APC and the validation dataset, the real world validation method solely relies on the APC counter data, thus completely ignoring the validations dataset. This means that, if the APC has counted 10 passengers entering a bus in a particular stop, it is assumed that there were 10 validations, even though the validations dataset may contain fewer entries, due to non-validating passengers. The downside is that due to the fact that only a subset of trips have APC data the data available for the real world validation is limited.

For the APC based validation data, each of the algorithms is used to predict an alighting station. Next, each estimated alighting stop is mapped to the district. After grouping the data on the district level, for each district of the city (where the bus with APC data passed for that day) the total number of passengers estimated to have exited a bus is calculated. This number is denoted as $estimated\_count\_out$. The $actual\_count\_out$ for each district is also extracted by aggregating the $count\_out$ field on the district level. The output of these steps is a compact dataset as shown below for each day and for each estimation algorithm:

```
district,counter_out,estimated_count_out
Annelinna,106,156
Ihaste,45,20
Karlova,210,199
Kesklinna,290,380
Supilinna,80,20
Tähtvere,77,55
Veeriku,103,81
```

From the above the relative error between the real and estimated count vectors can be calculated by applying the vector relative error formula described in 4.2.1.

However, as visible in the sample data above, relying purely on this relative error does not reflect the fact that there are varying number of $counter\_out$ by district. 290 passengers exited the bus in Kesklinna, while only 45 in Ihaste. A single misestimation of the district in Ihaste increases the relative error much more than the same misestimation in Annelinna. To tackle this a weighted relative error, which gives more weights to districts with more actual counter out is introduced.

Assuming the presence of APC count out information for k districts for the entire day, the counter out value can be denoted as D_i for district i, where $i \leq k$. The respective weights per each district would be $w_i = \frac{D_i}{\sum D}$, where $\sum D$ represents the sum of all D_i values. These weights w_i provide a relative measure of the contribution of each district to the total counter out values.

The weighted relative error is then calculated by using the formula described in equation 4.2.1 by using the same real and estimated vectors per district and the above weights w.

In addition to the calculation of the errors mentioned so far, a linear regression analysis was also performed. Given the expectation of equivalence between the $estimated\_count\_out$ and $actual\_count\_out$ values for each district, a linear regression was conducted to examine the relationship between these two variables across an extended time frame. The objective was to observe whether the regression line would tend towards the identity line, y=x, signifying a strong alignment between the estimated and actual counts.

For the linear regression analysis, actual count out and estimated count out for each district for 279 days of the year 2022 were obtained. These two extensive sets of data were then used to conduct the regression analysis.

4.3 Pipeline

This section provides a brief overview of the developed system architecture. The system consists of two main modules, namely the daily pipeline and the aggregator module as can be seen in Figure 5.

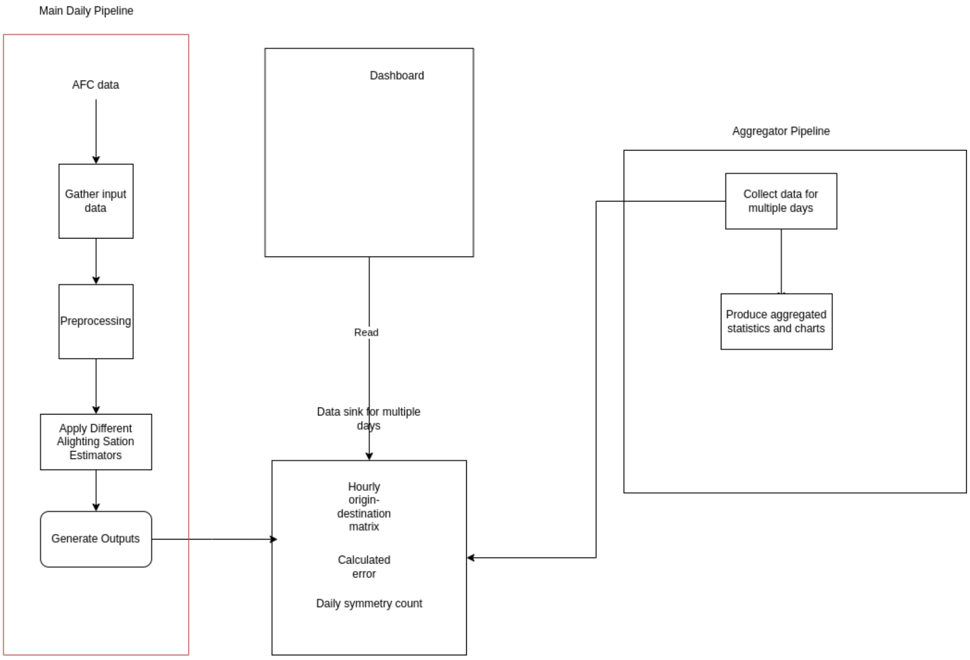

Main Daily Pipeline
AFC data
Gather input data
Preprocessing
Apply Different Alighting Sation Estimators
Generate Outputs
Dashboard
Read
Data sink for multiple days
Hourly origin-destination matrix
Calculated error
Daily symmetry count
Aggregator Pipeline
Collect data for multiple days
Produce aggregated statistics and charts

Figure 5. Modelling pipeline

The daily pipeline is responsible for downloading the AFC system data and generating the outputs for a single day. Outputs include district level OD matrix and district level errors for the day. On the other hand, the aggregator module takes the outputs generated by the daily pipeline for a certain number of days, denoted by n, and aggregates them to perform different analysis and error evaluation across multiple days.

Decoupling the main daily pipeline and the aggregator is important because it allows for the automatic scheduling of the main pipeline to run every day and store the output data, which can then be aggregated at a later time. The scheduling ability becomes important when the pipeline serves as a real-time data feed for the dashboard which is explained in 9.1. The separation between the two modules also ensures that the daily pipeline can run independently without being affected by the aggregation process, and that the aggregator can process data from multiple days at once which is useful for validation and charting.

The daily pipeline was entirely build using Python language. All the experiments were performed using a machine with below parameters:

```
CPU: AMD® Ryzen 7 4700u
OS: $Pop!_OS 20.04 LTS$
Ram: 16GB
Machine: Hp Envy x360
```

5 Experiments

This chapter presents the experiment-based evidence for the assumptions made by the algorithms and the experiments conducted to evaluate the performance of the alighting estimation algorithms. The results are reported using tables, figures, or graphs.

Firstly, some evidence about Assumption 2 which assumes the symmetry of passenger traffic on opposing routes is provided. It is followed by similar evidence about Assumption 3 which assumes the symmetry of passenger traffic on opposing routes and in relation to a midday cutoff time.

Next, the alighting estimation algorithms are validated using both synthetic and real-world validation methods.

Lastly, a comparison between the algorithms follows in the results interpretation section.

5.1 Evidence for assumptions of symmetry

This section aims to provide evidence to support two of the assumptions made in the methodology section, specifically in the context of the city of Tartu.

5.1.1 Bus trips symmetry in route level

To validate the assumption of symmetry, the daily validation dataset is utilized. As seen in section 3.1.1 this dataset includes information on the route, direction, timestamp, and validation location. By identifying opposing routes within the city, it is possible to compare the number of passengers traveling in each direction. This provides evidence that bus passengers in Tartu use the bus in a symmetrical way, meaning that a passenger traveling from stop A to stop B is likely to also travel from stop B to stop A. It is important to note that the validation of symmetry is subject to a certain degree of error and cannot be generalized to all cities and routes.

More concretely the below dataset presents 2 different trips following the same route but in different directions:

```
Trip Id RouteId Route Direction
123,     2,       A>B
134      2,       B>A
```

As a first step opposing trips for each route are identified. Then, the number of validations in each direction, i.e., towards the final destination and towards the opposite direction is counted.

This is an example of the result of the calculations for one route:

```
route_id: 371592
directions_start: ['A>B']
directions_opposite: ['B1>A', 'B>A']
route_long_name: FI - Nõlvaku
Route_Start_Count_Trips: 938
Route_Opposite_Count_Trips: 938
Route_Start_Count_validations: 12222
Route_Opposite_Count_validations: 12324
```

For the route FI-Nõlvaku, for one specific date, there have been exactly the same number of trips in one and the opposite direction. Across these trips there have been 12222 boardings in the first direction and 12324 boardings in the second direction.

By analyzing the number of boardings in each direction, it can be observed a certain level of symmetry between the route directions, for this specific example. This supposed symmetry can be quantified by calculating calculate the relative error between the number of boardings in each direction, expressed as a percentage of the total number of boardings on that route.

The relative error between boarding counts on opposing directions was computed for all routes in the year 2022, using daily validation data. Figure 6 shows the heatmap of relative distances for one week from September 19 to 25, 2022.

As can be seen from the heatmap in Figure 6, the relative error between boardings in opposing directions for all routes is below 5% during the selected week. This level of symmetry is consistent throughout the calendar week, including weekends, and is also observed throughout the year, indicating the general symmetry of bus usage in Tartu on route level.

5.1.2 Bus trips symmetry in route level from a cutoff time

Similarly to the above, some evidence for assumption of symmetry across a midday time is provided in this section (Assumption 3). The base around this assumption was that passengers generally take the bus from stop A to stop B in the morning and then take the bus again from B to A in the afternoon.

To test this hypothesis, the midday time of 14:00 in local time is selected as the midday point. After analyzing the boarding data for all routes, the directions of the same route are grouped into two groups, first and opposing directions. As in the previous section the number of boardings per each direction in the morning and afternoon is counted. A sample row of data for one route looks like below:

```
Route ID: 371604
Start directions: ['A>C>B', 'A>B1']
Opposite directions: ['B>C>A', 'B>A']
```

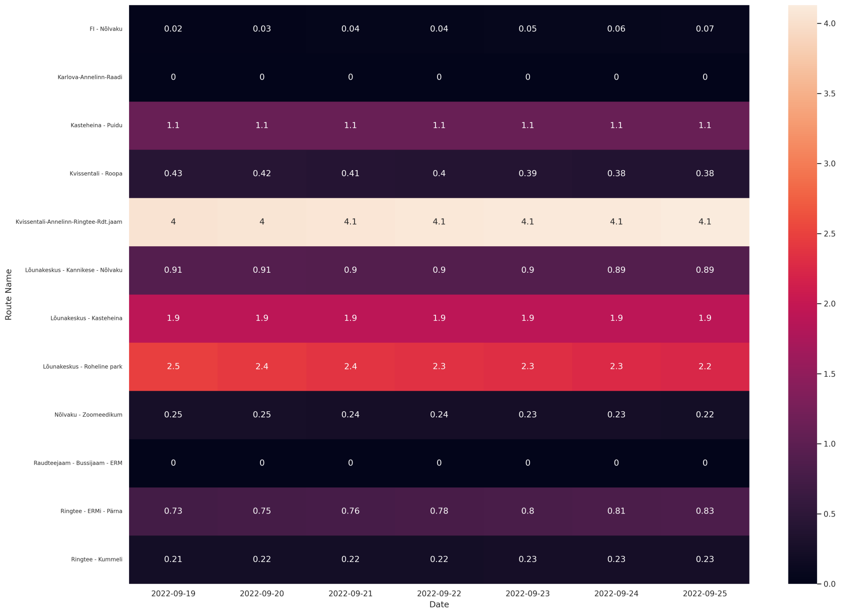


Figure 6. Heatmap showing the relative error between boardings in opposing directions as a percentage for each route for one week

```
Route long name: Lõunakeskus - Roheline park
Route start trip counts: 770
Route opposite trip counts: 756
Morning Route start count validations: 3058
Afternoon route start count validations: 3323
Morning route opposite count_validations: 2340
Afternoon route opposite count validations: 3546
```

After calculating the number of boardings in the morning or afternoon in the start or opposing direction, the relative error is calculated.

It is expected that the counts of Morning Route Start Count validations and Afternoon Route Opposite Count validations to match, and vice versa. The heatmaps in Figure 7 and Figure 8 show the results across one calendar week from March 21, 2022, to March 27, 2022.

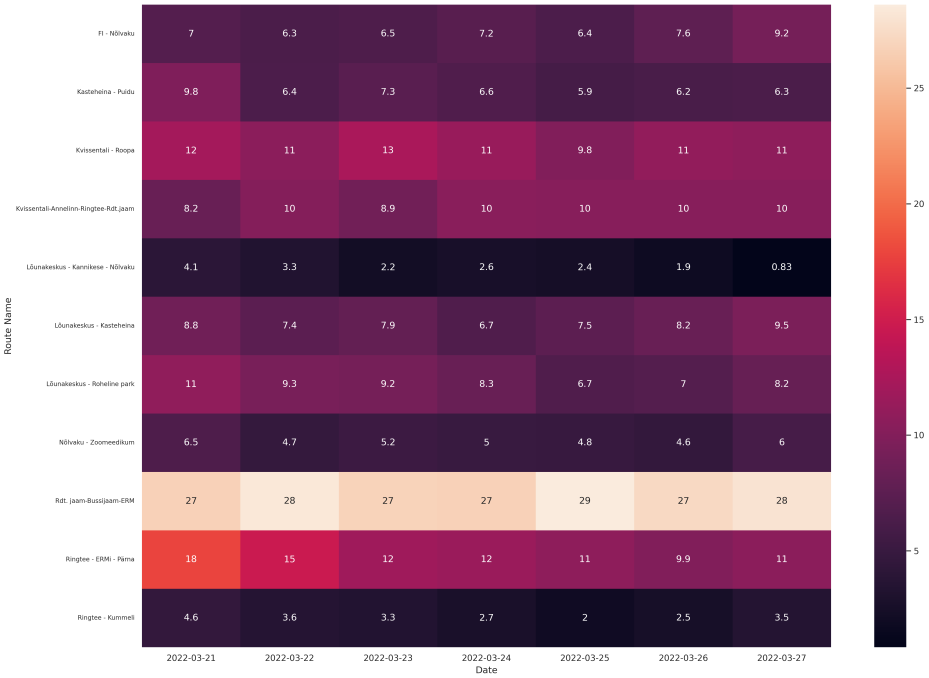


Figure 7. Heatmap illustrating the relative error between morning boardings count in first direction and evening boardings count in the opposing direction as a percentage for each route for one calendar week.

The relative error tends to remain consistently below 10% for most routes, with some outliers present. It is worth noting that the error may be slightly higher due to cutting the dataset in half for each count, making the relative error more sensitive. Moreover, it is observed that the error does not show significant variation between weekdays and weekends and is consistent throughout the year.

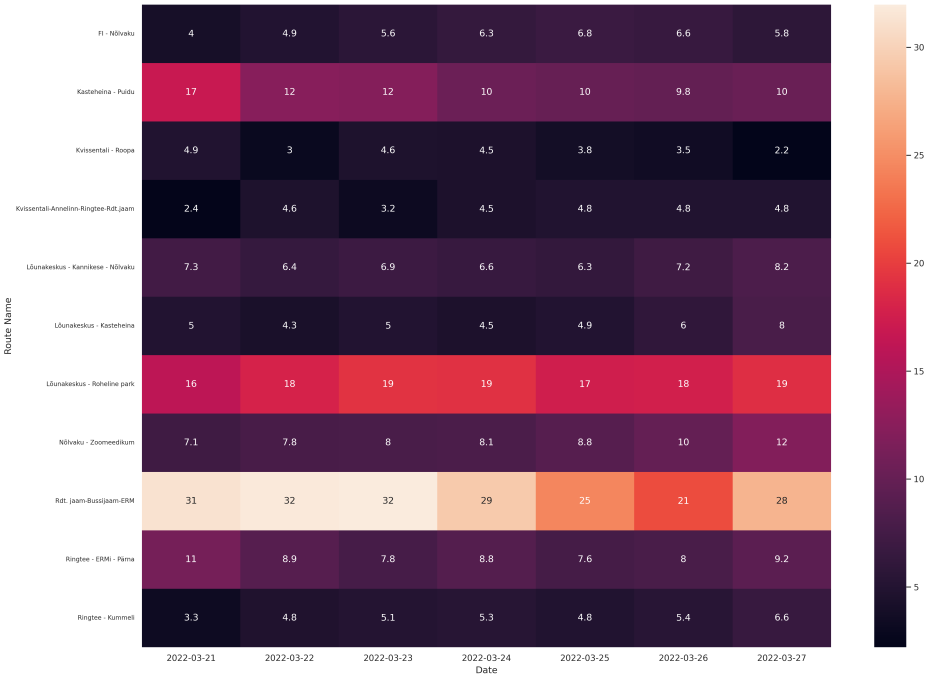


Figure 8. Heatmap illustrating the relative error between evening boardings count in first direction and morning boardings count in the opposing direction as a percentage for each route for one calendar week.

5.2 Alighting estimation methods evaluation and comparison

This section of the experiments aims to show the results of the synthetic and real world validation, after applying the four alighting estimation methods for a number of days in the year 2022.

5.2.1 Synthetic validation method

For the synthetic validation method the district level relative error was calculated for all 365 days of the year to evaluate the performance of each of the methods.

Table 2 shows the comparison between the 4 different alighting methods by using the synthetic validation method as explained in section 4.2.2. The table displays the mean, standard deviation, minimum, maximum, and median values of the calculated relative error as explained in the methodology section. To provide a visual representation of the

Table 2. Comparison of the synthetic evaluation metric statistics across 365 days for the 4 methods.

	Results				
	Mean	**Std deviation**	**Min**	**Max**	**Median**
Uniform distribution	36.6%	27.5%	24.7%	47.7%	36.5%
Normal distribution	26.28%	25.8%	14.6%	36.2%	26.2%
Symmetric estimation	14.95%	16.67%	10.76%	21.36%	14.72%
Midday Symmetric estimation	14.54%	18.00%	10.51%	24.54%	14.24%

distribution of the evaluation metric, Figure 9 displays the same information as a box plot.

The results show that the normal distribution method had a relative error of 26.28% while the uniform distribution displays the highest error of 36.6%. The symmetric estimation and midday symmetric estimation methods show improvements by having similar mean values, both below 15%. Additionally, the standard deviation is the highest for the uniform distribution method (27.5%) and lowest for the symmetric estimation method (16.67%).

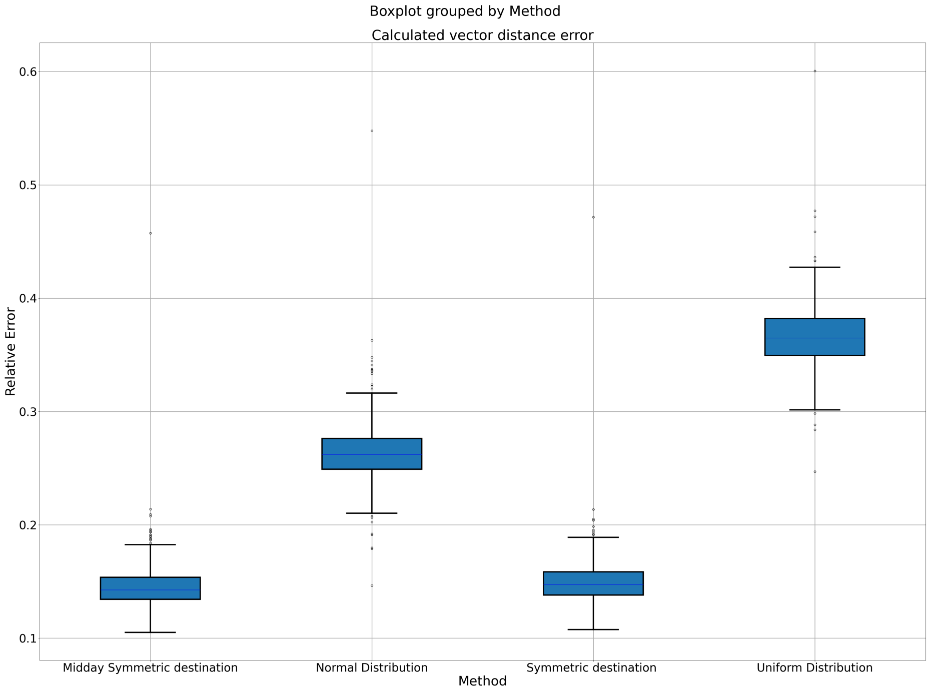


Figure 9. Boxplot showing variance of the synthetic evaluation metric across 365 days.

In the boxplot in Figure 9 it is noticeable that the uniform distribution and normal distribution method have a larger spread and higher range of outliers on both directions. The spread becomes much lower in Symmetric estimation method and presents itself with outliers only in one direction.

The bar chart depicted in Figure 10 illustrates the mean values of the output of the synthetic validation method across 365 days of the year, as computed for each method. This visualization provides a clear indication of the relative performance of the methods in terms of their average accuracy.

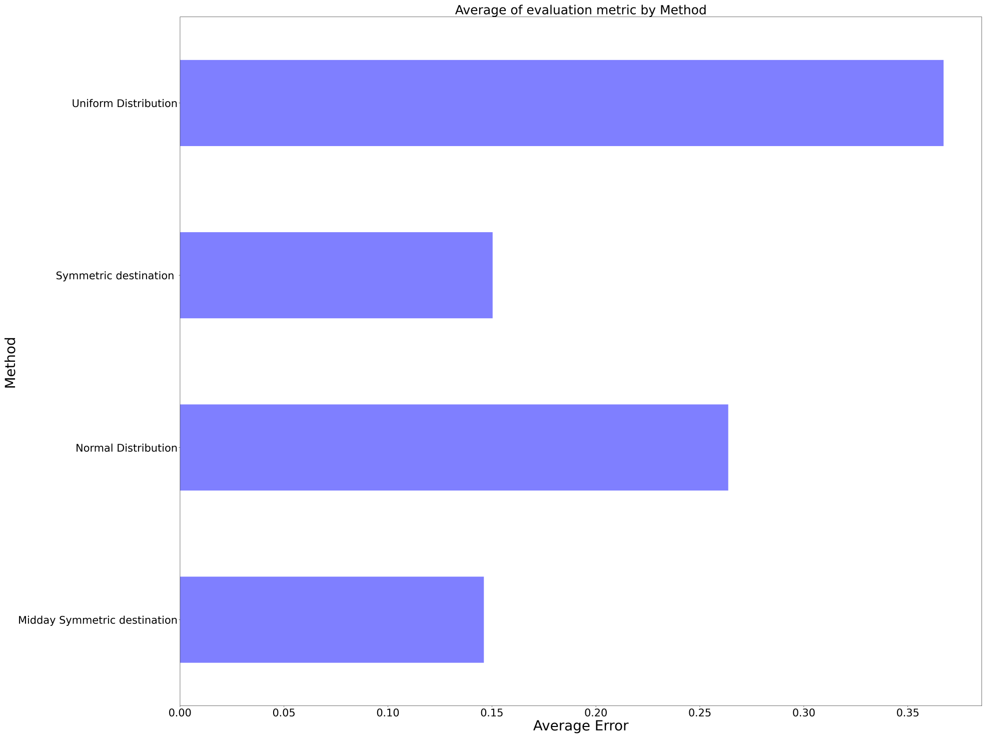


Figure 10. Bar chart displaying the mean of the evaluation metric across 365 days

For additional insights, Figure 11 shows a line chart for all the calculated data points across 365 dates. The x axis represents each date and the y axis the calculated evaluation metric for each method.

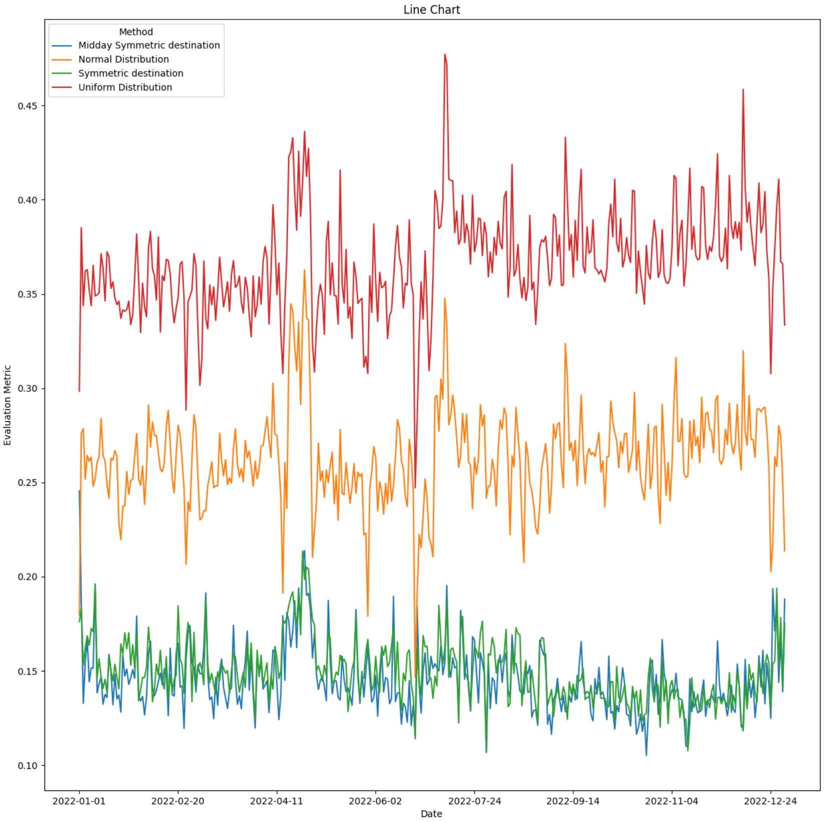


Figure 11. Line chart illustrating the values of the sythentic validation metric across 365 days for all the methods.

5.2.2 Counter based validation method

After executing the steps described in the methodology section 4.2.3 results are presented below both when using the relative error and weighted relative error in district level.Table 3 shows statistics of the district level non-weighted relative error for each day of the 2022 year which has more than 300 counts from APC data per day. Similarly, table 4 shows the weighted error.

RMSE and normalized RMSE metrics are displayed on table 5.

Table 3. Comparison of the district level relative error across 279 days for 4 methods

	Results				
	Mean	**Std deviation**	**Min**	**Max**	**Median**
Uniform distribution	41.51%	8.6%	18.21%	63.87%	42.05%
Normal distribution	41.22%	11.32%	12.78%	79.19%	40.83%
Symmetric estimation	32.63%	11.48%	10.49%	71.78%	31.56%
Midday Symmetric est.	38.14%	17.18%	11.93%	130.03%	35.46%

Table 4. Comparison of the district level weighted relative error across 279 days for 4 methods

	Results				
	Mean	**Std deviation**	**Min**	**Max**	**Median**
Uniform distribution	40.46%	9.96%	12.06%	68.25%	40.75%
Normal distribution	37.48%	13.33%	7.00%	79.66%	37.13%
Symmetric estimation	28.45%	11.44%	5.52%	68.73%	28.03%
Midday Symmetric est.	33.10%	17.34%	7.77%	134.14%	30.14%

Table 5. Comparison of the RMSE and Normalized RMSE for 4 methods

	Results	
	RMSE	**Normalized RMSE**
Uniform distribution	31.62	0.68
Normal distribution	32.30	0.67
Symmetric estimation	24.49	0.53
Midday Symmetric est.	28.46	0.61

Figures 12 and 13 illustrate the same as a box plots while Figure 14 shows a bar chart of the mean of the calculated errors for each method.

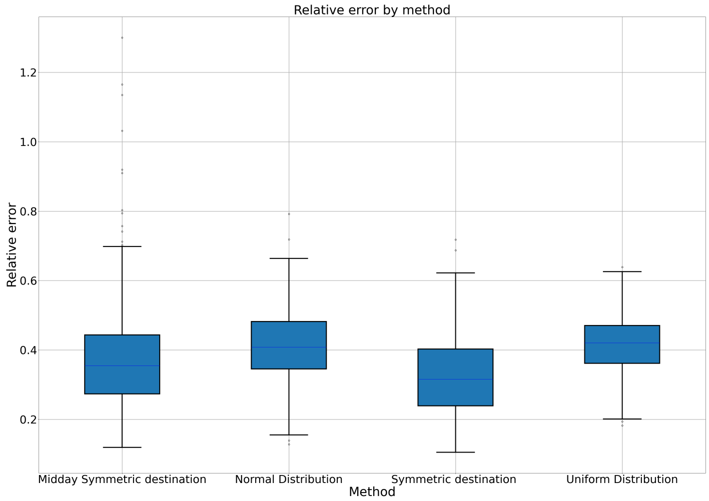


Figure 12. Boxplot showing variance of the district level relative error across 279 days of the year 2022.

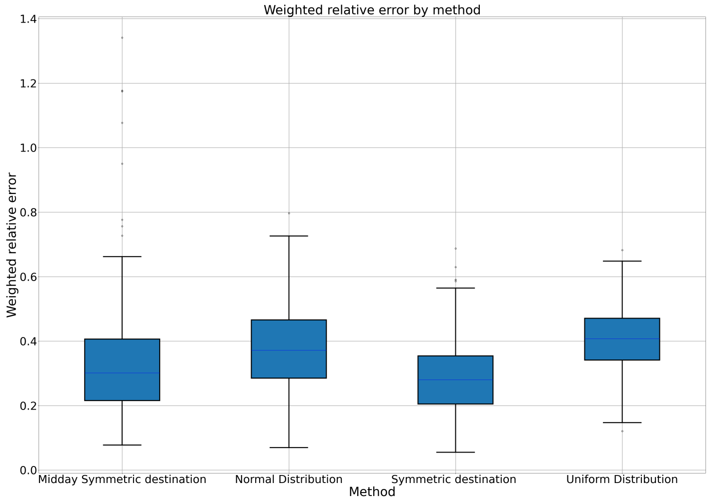


Figure 13. Boxplot showing variance of the district level weighted relative error across 279 days of the year 2022.

Lastly the linear regression approximation between count out and estimated count out per district per 279 days are displayed in figure 15.

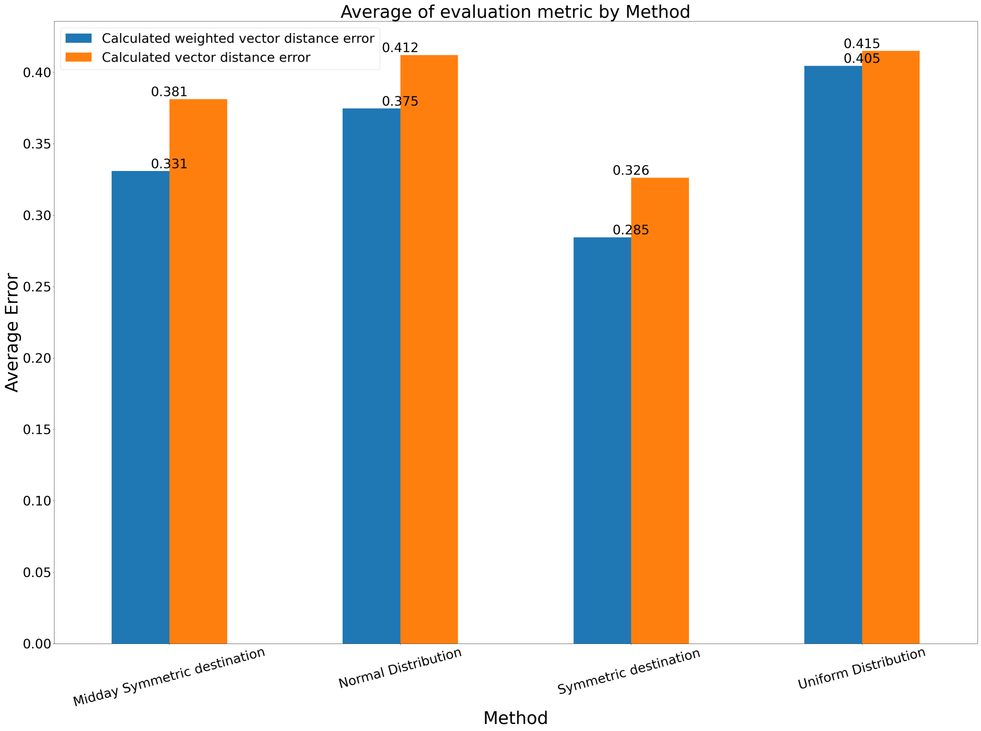
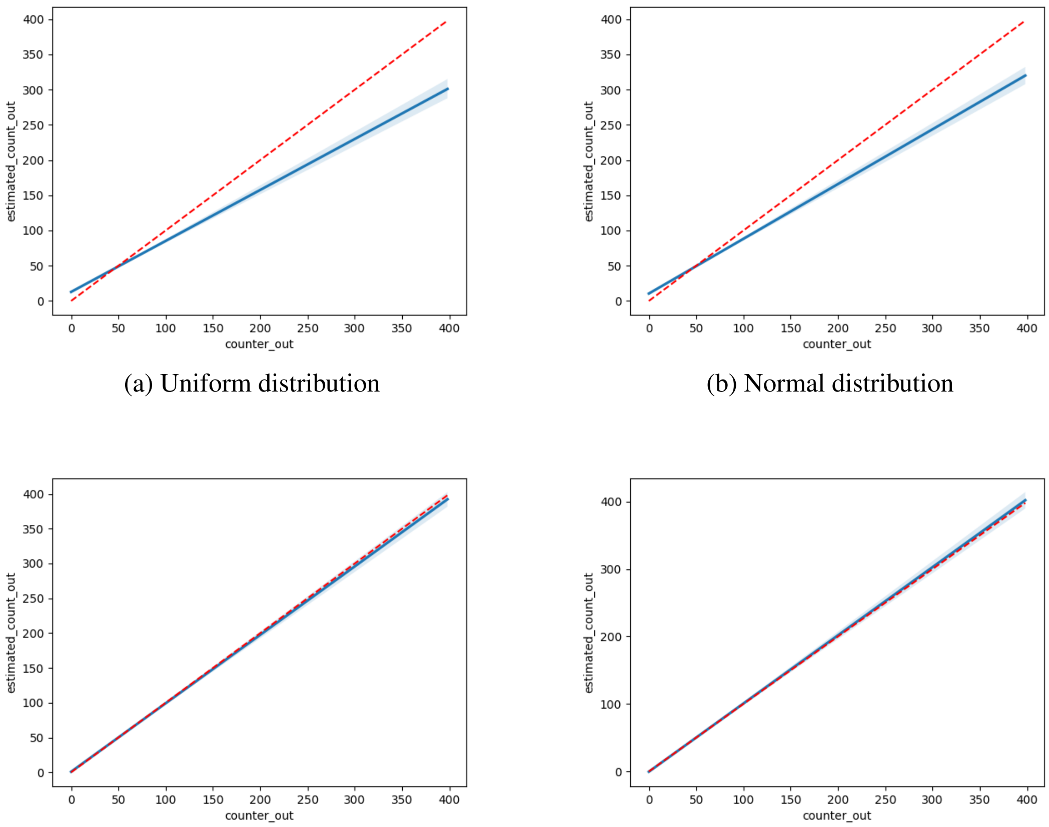


Figure 14. Bar chart showing means of district level weighted relative error across 279 days of the year 2022 for the four methods.

5.3 Interpreting the results

Observing the outputs of the experiments, some conclusions about the performance of various methods can be drawn. As it was anticipated due to the simplicity of the method, the uniform distribution selector was the worst performer. Randomly picking the end station resulted in a high error and both the synthetic and real world validation methods picture the same story. On the other hand, the normal distribution method showed a significant improvement as per the evaluation metrics, despite being based on a simple assumption and having a simple implementation. According to the analysis, the normal distribution ranks better than uniform distribution in all the error calculations besides the RMSE where the error is slightly higher. It is to be noted the quite big improvement in this method when checking in the weighted real world relative error versus the non-weighted one. Linear regression line also paints similar picture where the regression line for the estimated counts per district tends to be closer to the y=x line than uniform distribution.

According to every measurement, the symmetric and midday symmetric estimation



(a) Uniform distribution

(b) Normal distribution

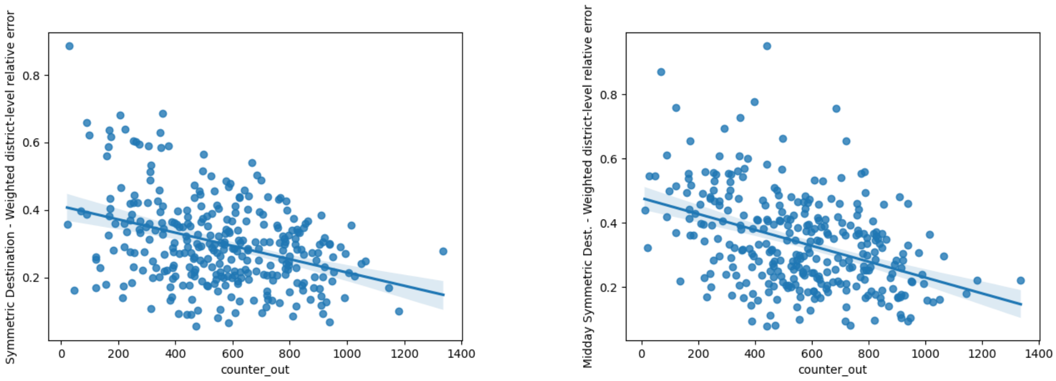
(c) Symmetric estimation

(d) Midday Symmetric estimation

Figure 15. Linear regression approximation

methods showed significant improvement compared to the two simpler methods discussed above. Deciding the most accurate between the two, however, needs discussion as the synthetic validation method and real world validation method do not paint the same picture. Synthetic validation method metric, shows the midday symmetric estimation method to have a slightly lower error of 14.54% against the 14.95% of the symmetric estimation method. However, the symmetric estimation showed less volatility in the error values by having a smaller standard deviation. On the other side, the real world validation method based on APC data shows the symmetric estimation method to have the lowest error by a considerable margin; 5.51% and 4.65 % respectively between the non-weighted version of the district level relative error. Results retrieved from RMSE and NRMSE show the same view. The linear regression line for both of the methods is very close to the y=x line, displaying encouraging results.

One particularly interesting thing to note from the results of the real world validation method is the comparison between the weighted and non-weighted versions of the district relative error metric. Every method displayed a lower error in the weighted version. The midday symmetric method had a weighted error of 5.04% lower, followed by the symmetric estimation with 4.15%, normal distribution with 3.74% and uniform distribution with 1.05% lower. Firstly, this affirms the previous statement that midday symmetric and symmetric estimation methods perform better overall, shown by the highest gain when taking into account the big difference of counts between districts. Secondly this shows that the relative error is smaller and the estimation's accuracy is higher when there are more APC data. Considering that according to the EDA performed in section 3.2.2, for the 2022 year only 1.75% of trips per day were covered with APC data, the calculated relative errors could have shown improvement had more APC data been available. This same is also shown in Figure 16, which proves that there is a negative correlation between the number of count out per day and the weighted relative error.



(a) Symmetric estimation

(b) Midday Symmetric estimation

Figure 16. Scatter plot displaying negative correlation between number of count out and calculated relative district level error.

6 Discussion

This thesis addressed the issue of missing alighting information in Tartu's bus AFC system data by proposing four different probability based alighting estimation methods. The methods were verified using synthetic and real-world validation methods. Results showed that two of the proposed algorithms outperformed the rest in terms of accuracy and error. The thesis also demonstrated the value of having complete passenger data, providing examples of insights that can be derived from analyzing both boarding and alighting information, such as origin-destination matrices.

Both the proposed methods can be used for day to day monitoring of origin destination matrices. Furthermore the validation methods and evaluation metrics can be used in future work of the city of Tartu as a baseline metric in lack of abundant real world information about alighting stations.

6.1 Limitations

This study is accompanied by several limitations that should be acknowledged in order to comprehend the scope and potential impact of the findings accurately:

- The real-world validation method relies on data from APC, which often only capture a minor fraction of the overall passenger trips. While efforts were made to mitigate this limitation through the application of relative error and weighted relative error metrics, it remains crucial to remain aware of this constraint.

- The validation level estimation in individual trip level, because of the probabilistic methods used, can be inaccurate, making it unsuitable for determining station-to-station origin-destination matrices.

- The probabilistic methods used in the study are underpinned by several assumptions that may not universally hold true across all scenarios in the real world.

- While efforts were made to establish evidence of symmetry in boardings in the route level, it is important to acknowledge that further insights are required to conclusively affirm this assumption.

- Due to the four probabilistic methods being only applied in the context of the city of Tartu, drawing conclusions about the degree of their generalizability to diverse urban settings and countries with distinct infrastructures and passenger behaviors is unfeasible.

These limitations highlight the areas in which the study may have inherent constraints or uncertainties. While they influence the extent to which the presented findings can be generalized, they also suggest avenues for future research to refine and expand upon the current methodology and insights.

6.2 Future work

In terms of future work, there are several avenues that can be explored.

Firstly, if possible, it would be beneficial to include masked passenger identifier information in the data source. This could potentially enable the implementation of a trip chaining model algorithm, which could improve the accuracy of trip-level alighting information and aid in the creation of an origin-destination matrix at the station level. The evaluation metrics in the district-level calculated in this study would serve as a benchmark.

Secondly, an avenue for future exploration involves the utilization of more advanced probabilistic techniques. While the methods outlined in this study constitute an initial endeavor in estimating alighting stations for Tartu city by incorporating passenger boarding data, there exists the opportunity to incorporate additional information for enhancing station prediction accuracy. For instance, integrating data such as population density records or even land-use patterns could potentially yield more comprehensive and refined outcomes. This extension could contribute to a more nuanced understanding of station dynamics and passenger behaviors.

Thirdly, while this thesis focused on probabilistic methods for destination estimation, there is an opportunity to explore the use of more sophisticated deep learning approaches. With recent developments in artificial intelligence and the growing availability of data, it may be possible to improve the accuracy of destination estimation for the city of Tartu using deep learning models. This could involve exploring various neural network architectures and training methods, and may require a larger and more diverse dataset than the one used in this work. Additionally, it may be possible to integrate deep learning methods with the trip chaining model algorithm suggested in the previous point, further improving the accuracy of origin-destination matrix estimation.

Next, another promising avenue for exploration revolves around delving into the generalizability of the probability methods used in this study, extending their application beyond the current context. The investigation of how these methods perform in diverse countries and cities holds significant potential for uncovering valuable insights not only about the broader applicability of the methods, but also on the differences, if any, of passenger behaviour in different cities.

Lastly, while the assumption made about the symmetry are reasonable, additional insights into passenger behavioral patterns would be valuable. This could provide a better understanding of public transport passengers' behavior, not just in Tartu but more generally, and could be used as a starting point for future work.

7 Summary

Entry-only AFC systems, such as the one implemented in Tartu bus transportation system are a common source of trying to understand human mobility patterns. These systems contain accurate passenger-level data regarding boarding station and time, however due to their working, they fall short on obtaining information about alighting station and time. To allow extraction of OD-matrices which enable good understanding of passenger mobility flows, these data are necessary. A considerable amount of work has been done by researchers to address this limitation of entry-only AFC systems through the use of different alighting station estimation methods, which can be grouped into three groups of models; the trip chaining models, probability models and deep learning models.

The primary objective of the thesis was to estimate the alighting information for the city of Tartu while building on top of existing research. Through the utilization of four distinct probabilistic methods, varying in complexity, this study was able to accomplish this estimation task. The methodologies underwent two validation methods, involving both synthetic and real-world validation processes encompassing a year's worth of data. These approaches were then subjected to a comprehensive comparative analysis.

Of the employed methods, the symmetric destination estimation and midday symmetric destination estimation techniques drew upon the established concept that human mobility patterns often demonstrate regularity and symmetry. Consequently, evidence supporting the presence of such patterns within the confines of Tartu city was gathered and presented.

Within the suite of four methods employed, two methodologies emerged as particularly adept: the symmetric destination estimation method and the midday symmetric destination estimation method. Notably, these approaches exhibited higher accuracy in alighting station estimation at the district level, surpassing the performance of their simpler counterparts—the uniform distribution and the normal distribution models.

In essence, this thesis contributes by bridging the gap between entry-only AFC data and alighting information in the context of the city of Tartu, hence enhancing the grasp of urban mobility patterns. The combination of probabilistic estimation techniques, validation procedures, and the real world application of these methods advance the field, thereby presenting a good foundation for future research and practical applications. Simultaneously, the outcomes stand to offer valuable advantages to Tartu's bus transportation planners, enhancing their decision-making processes.

8 Conclusion

This thesis aimed to extract alighting information for entry-only AFC system in the city of Tartu. The task was achieved through the use of four probabilistic models. All the models were validated using a synthetic validation method, which relied only on boarding information from AFC data during validation and a real-world validation method which relied on the data from APC present on a subset of the trips for each day. The two proposed symmetric estimation methods achieved the lowest weighted relative error on the district level for the real-world validation: 28.45% for the symmetric estimation method and 33.10% for the Midday symmetric estimation method. From the other two simpler methods, the same error was 40.46% for the uniform distribution and 37.48% for the normal distribution. The small APC data available could impact the validation. According to the gathered evidence, the error tends to grow smaller with the increase in APC data. Overall, this work provides a foundation for future research in developing more accurate and effective methods for estimating alighting information in public transport systems.

9 Publications

A paper based on the work presented in this thesis was published in Sensors journal [10]. In this paper, we designed a real-time system which used IoT devices installed in the city as a source of passenger, bike, bus and vehicle data. The built platform through different optimizations was able to calculate the split between usage of the mentioned modes of transportation on a daily time frame. Furthermore, the daily and hourly district-level OD matrixes for each transportation were built. These OD matrixes can be viewed and downloaded in the live dashboard.

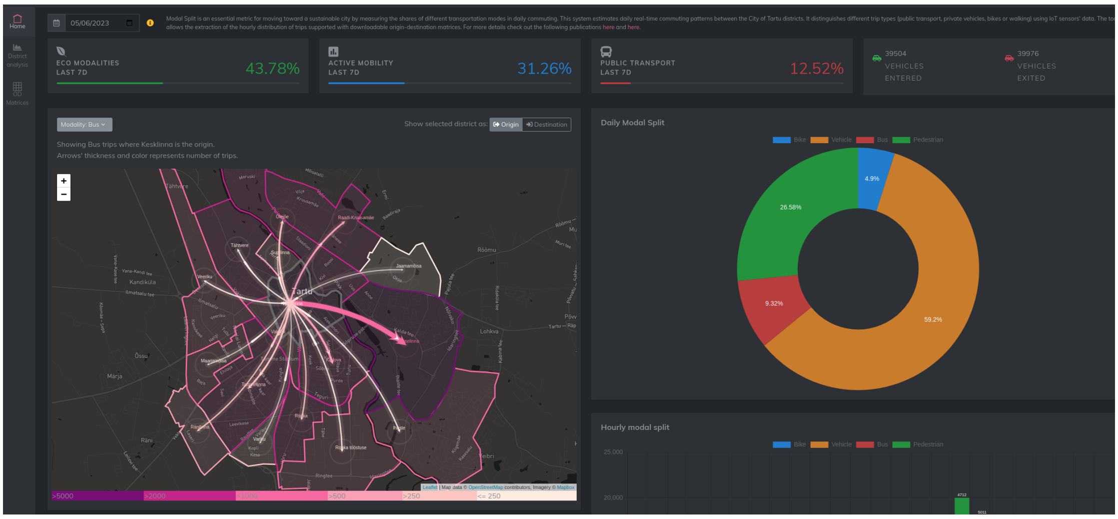
The paper, in its bus related work, uses several aspects of the work presented in this master's thesis. More specifically it uses the main daily pipepine and aggregator modules mentioned in Section 5 for processing of the bus data from the moment input files are retrieved up until the hourly and daily OD matrix generation. The alighting estimation method selected in the scope of the paper is the Midday Symmetric Estimation method described in Section 4.1.6.

Furthermore, I also contributed in co-building the referenced dashboard in displaying the output of the system. Some insights about the dashboard and the nature of my contribution are presented in the next section.

9.1 Dashboard

The dashboard was co-created as a website using various technologies, such as NodeJS, PostgreSQL, Javascript, and LeafletJs, to display the hourly district-level origin-destination matrix, among other things. In the context of this thesis, the dashboard serves as a real world example of some of the insights that can be enabled by having alighting information. The dashboard shows processed bus AFC system data for each day. The information presented on the dashboard includes the total number of bus passengers, the daily district-level OD matrix in the map section (see Figure 17), and the hourly district-level origin-destination matrix in the OD matrices menu (see Figure 18).

Both daily and hourly origin-destination matrices use the symmetric destination station estimation algorithm to determine the destination district. The dashboard is accessible using the link: `https://its.cs.ut.ee/modsplit/`. Each day the dashboard displays the OD matrices information for the previous day.



Modal Split is an essential metric for moving toward a sustainable city by measuring the shares of different transportation modes in daily commuting. This system estimates daily real-time commuting patterns between the City of Tartu districts. It distinguishes different trip types (public transport, private vehicles, bikes or walking) using IoT sensors' data. The to
allows the extraction of the hourly distribution of trips supported with downloadable origin-destination matrices. For more details check out the following publications here and here.

ECO MODALITIES LAST 7D 43.78%

ACTIVE MOBILITY LAST 7D 31.26%

PUBLIC TRANSPORT LAST 7D 12.52%

39504 VEHICLES ENTERED

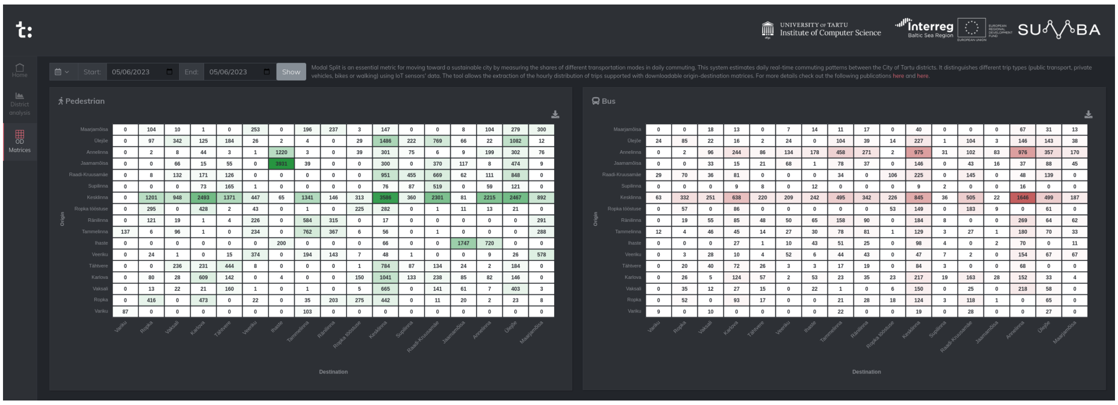
39976 VEHICLES EXITED

Modality: Bus

Show selected district as: Origin Destination

Showing Bus trips where Kesklinna is the origin.
Arrows' thickness and color represents number of trips.

Figure 17. Dashboard home page



Modal Split is an essential metric for moving toward a sustainable city by measuring the shares of different transportation modes in daily commuting. This system estimates daily real-time commuting patterns between the City of Tartu districts. It distinguishes different trip types (public transport, private vehicles, bikes or walking) using IoT sensors' data. The tool allows the extraction of the hourly distribution of trips supported with downloadable origin-destination matrices. For more details check out the following publications here and here.

Pedestrian

Origin	Variku	Ropka	Vaksali	Karlova	Tähtvere	Veeriku	Ihaste	Tammelinna	Ränilinna	Ropka tööstuse	Kesklinna	Supilinna	Raadi-Kruusamäe	Jaamamõisa	Annelinna	Ülejõe	Maarjamõisa
Maarjamõisa	0	104	10	1	0	253	0	196	237	3	147	0	0	8	104	279	300
Ülejõe	0	97	342	125	184	26	2	4	0	29	1486	222	769	66	22	1082	12
Annelinna	0	2	8	44	3	1	1220	3	0	39	301	75	6	9	199	302	76
Jaamamõisa	0	0	66	15	55	0	3931	39	0	0	300	0	370	117	8	474	9
Raadi-Kruusamäe	0	8	132	171	126	0	0	0	0	0	951	455	669	62	111	848	0
Supilinna	0	0	0	73	165	1	0	0	0	0	76	87	519	0	59	121	0
Kesklinna	0	1201	948	2493	1371	447	65	1341	146	313	3586	360	2301	81	2215	2467	892
Ropka tööstuse	0	295	0	428	2	43	0	1	0	225	282	0	1	11	13	21	0
Ränilinna	0	121	19	1	4	226	0	584	315	0	17	0	0	0	0	0	291
Tammelinna	137	6	96	1	0	234	0	762	367	6	56	0	1	0	0	0	288
Ihaste	0	0	0	0	0	0	200	0	0	0	66	0	0	1747	720	0	0
Veeriku	0	24	1	0	15	374	0	194	143	7	48	1	0	0	9	26	578
Tähtvere	0	0	236	231	444	8	0	0	0	1	784	87	134	24	2	184	0
Karlova	0	80	28	609	142	0	4	0	0	150	1041	133	238	85	82	146	0
Vaksali	0	13	22	21	160	1	0	1	0	5	665	0	141	61	7	403	3
Ropka	0	416	0	473	0	22	0	35	203	275	442	0	11	20	2	23	8
Variku	87	0	0	0	0	0	0	103	0	0	0	0	0	0	0	0	0

Destination

Bus

Origin	Variku	Ropka	Vaksali	Karlova	Tähtvere	Veeriku	Ihaste	Tammelinna	Ränilinna	Ropka tööstuse	Kesklinna	Supilinna	Raadi-Kruusamäe	Jaamamõisa	Annelinna	Ülejõe	Maarjamõisa
Maarjamõisa	0	0	18	13	0	7	14	11	17	0	40	0	0	0	67	31	13
Ülejõe	24	85	22	16	2	24	0	104	39	14	227	1	104	3	146	143	38
Annelinna	0	2	96	244	86	134	178	458	271	2	975	31	102	83	976	357	170
Jaamamõisa	0	0	33	15	21	68	1	78	37	0	146	0	43	16	37	88	45
Raadi-Kruusamäe	29	70	36	81	0	0	0	34	0	106	225	0	145	0	48	139	0
Supilinna	0	0	0	9	8	0	12	0	0	0	9	2	0	0	16	0	0
Kesklinna	63	332	251	638	220	209	242	495	342	226	845	36	505	22	1646	499	187
Ropka tööstuse	0	57	0	86	0	0	0	0	0	53	149	0	183	9	0	61	0
Ränilinna	0	19	55	85	48	50	65	158	90	0	184	8	0	0	269	64	62
Tammelinna	12	4	46	45	14	27	30	78	81	1	129	3	27	1	180	70	33
Ihaste	0	0	0	27	1	10	43	51	25	0	98	4	0	2	70	0	11
Veeriku	0	3	28	10	4	52	6	44	43	0	47	7	2	0	154	67	67
Tähtvere	0	20	40	72	26	3	3	17	19	0	84	3	0	0	68	0	0
Karlova	0	26	5	124	57	2	53	23	35	23	217	19	163	28	152	33	4
Vaksali	0	35	12	27	15	0	22	1	0	6	150	0	25	0	218	58	0
Ropka	0	52	0	93	17	0	0	21	28	18	124	3	118	1	0	65	0
Variku	9	0	10	0	0	0	0	22	0	0	19	0	28	0	0	27	0

Destination

Figure 18. Dashboard bus hourly origin-destination matrix

References

[1] Peatus EE. `http://peatus.ee/gtfs/`. Accessed Aug 5, 2023.

[2] Tartu Bus Card. `https://www.tartu.ee/en/tartu-bus-card`. Accessed May 7, 2023.

[3] Tartu Pilet. `https://tartu.pilet.ee/buy`. Accessed Aug 5, 2023.

[4] Azalden A Alsger, Mahmoud Mesbah, Luis Ferreira, and Hamid Safi. Use of smart card fare data to estimate public transport origin–destination matrix. *Transportation Research Record*, 2535(1):88–96, 2015.

[5] James J Barry, Robert Newhouser, Adam Rahbee, and Shermeen Sayeda. Origin and destination estimation in new york city with automated fare system data. *Transportation Research Record*, 1817(1):183–187, 2002.

[6] H. Dou, H. Liu, and X. Yang. Od matrix estimation method of public transportation flow based on passenger boarding and alighting. 25:79–82, 2007.

[7] Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. Understanding individual human mobility patterns. *nature*, 453(7196):779–782, 2008.

[8] Y.U. Jie and X.G. Yang. Estimation a transit route od matrix using on/off data: An application of modified bp artificial neural network. *Systems Engineering*, 24(1):89–92, 2006.

[9] J. Jung and K. Sohn. Deep-learning architecture to forecast destinations of bus passengers from entry-only smart-card data. *IET Intelligent Transport Systems*, 11(1):334–339, 2017.

[10] Kaveh Khoshkhah, Mozhgan Pourmoradnasseri, Amnir Hadachi, Helen Tera, Jakob Mass, Erald Keshi, and Shan Wu. Real-time system for daily modal split estimation and OD matrices generation using IoT data: A case study of Tartu city. *Sensors*, 22(8), 2022.

[11] Tian Li, Dazhi Sun, Peng Jing, and Kaixi Yang. Smart card data mining of public transport destination: A literature review. *Information*, 9(1):18, 2018.

[12] David S Navick and Peter G Furth. Estimating passenger miles, origin–destination patterns, and loads with location-stamped farebox data. *Transportation research record*, 1799(1):107–113, 2002.

[13] C.M. Schneider, V. Belik, T. Couronné, Z. Smoreda, and M.C. González. Unravelling daily human mobility motifs. *J. R. Soc. Interface*, 10(20130246), 2013.

[14] W. Yang, H. Wang, X. Ye, C. Xu, and Jiang. OD matrix inference for urban public transportation trip based on gps and ic card data. *Journal of Chongqing Jiaotong Universitu*, 34:117–121, 2015.

[15] M. Zhang, Y. Guo, and Y. Ma. A probability model of transit OD distribution based on the allure of bus station. *Journal of Transport Information and Safety*, 32:57–61, 2014.

[16] Juanjuan Zhao, Qiang Qu, Fan Zhang, Chengzhong Xu, and Siyuan Liu. Spatio-temporal analysis of passenger travel patterns in massive smart card data. *IEEE Transactions on Intelligent Transportation Systems*, 18(11):3135–3146, 2017.

[17] X. Zhou, X. Yang, and X. . J. Wu. OD matrix estimation method of public transportation flow based on passenger boarding and alighting. 40:1027–1030, 2012.

Appendix

I. Licence

Non-exclusive licence to reproduce thesis and make thesis public

I, **Erald Keshi**,
*(*author's name)

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to

   reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright,

   Alighting Estimation in Entry-Only AFC Systems; a Case Study of Tartu City,
   *(*title of thesis)

   supervised by Mozhgan Pourmoradnasseri and Amnir Hadachi.
   *(*supervisor's name)

2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.

4. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Erald Keshi
10/08/2023