USE OF MOBILE PHONE DATA FOR ANALYSIS OF NUMBER OF TRAIN TRAVELLERS

ABSTRACT

Several studies have pointed to the difficulties of obtaining good data on train ridership. There are at least two challenges regarding these data. First, train operators consider such data confidential business information, especially in high resolution. Second, the data that actually are available vary in quality and coverage. This paper studies mobile phone data as an alternative measure to obtain data about train ridership.

Handset counts were obtained from one telecom operator for selected mobile phone base stations and compared with timetable data and APC. The selected base stations are located so that it is likely that a large share of the mobile phone traffic is generated by train passengers. The number of units connected to a base station is found to correspond relatively well with the trains that pass close to the base stations. A ratio between the handset count and APC data appear as promising in utilising handset count to calculate train ridership, with ratios around one in the rush hours. We discuss preliminary results as well as methodological and technical challenges.

To make sure that we do not violate privacy concerns, the data used in the study have been approved by personal privacy representatives.

Keywords

Railway; Ridership; Train; Mobile phone data;

1 INTRODUCTION

1.1 Research on train ridership is important

When analysing public transportation, including trains, ridership is an important factor. The number of travellers is a measure of demand for transportation services, which is important information for planning and evaluations. With updated ridership information, planners should be able to get a detailed, continuous and accurate vision of the travel behaviour of their customers. This is important in planning and improving the transportation service. Other uses of ridership numbers are calibration and validation of transport models. Boyle (1998) identifies four main reasons why ridership data are collected. Firstly, ridership is reported to external funding and oversight agencies. Secondly, it monitors trends over time. Thirdly, ridership is a key performance indicator at various levels of the transportation system. Finally, ridership data identifies locations with the greatest boarding and alighting activity, which is important not only for its own purpose, but because the safe management of the railway may depend upon it. Other issues that call for data on ridership on trains include fare equipment location optimization, fare policy change and train schedule (Li, 2000). In addition, revenue distribution in integrated public transportation systems can be based on ridership data.

According to Vuchic (2005), the purpose of obtaining data on passenger volume and load count is to monitor trends and travel behaviour over time. Such data show passenger volumes on different sections of a line, the maximum number of passengers on different lines and when the maximum is reached, along with information on variations in passenger volume. Ridership data are also a key performance indicator in public transport (Vuchic, 2005).

On the other hand, the process of obtaining data on ridership creates least two major challenges. Firstly, train operators consider such data confidential business information, especially in high resolution (Vigren, 2017). Secondly, the data that actually are available vary in quality and coverage. Several studies highlight the unreliability of ridership data (including Chu and Chapleau, 2008 and Fowkes et al., 1985).

1.2 Mobile phone data and alternative technologies

Doi and Allen (1986) studied a rapid transit line for a period of about six and a half years (from 1978–1984) based on ridership data provided by a transit authority. Even though some studies combine several data sets, most studies on ridership rely on one data source. Wang et al. (2011) explored the application of archived data from Automated Data Collection Systems (ADCS) to transport planning with a focus on bus passengers' travel behaviour. They claimed that it was the first known attempt to validate the results by comparing automated ridership data with manual passenger survey data. Passenger distribution in the urban Copenhagen rail network is, according to Nielsen et al. (2014), tracked based on a combination of Electronic Weighing Equipment (EWE) and Automatic Passenger Count (APC). The two systems provide complementary information, since the weight-based estimation provides information about the total traffic volume and automatic passenger counting provides information on passenger flow. The two systems can also be used to perform quality assurance of each other’s measurements. Zhao et al. (2007) combine data from the automated fare collection system and the automated vehicle location system to examine the rail-to-bus trip sequence to obtain a clearer picture of ridership patterns. De Regt et al. (2017) combine smart card and Global System for Mobile Communications (GSM) data to examine spatial and temporal patterns of public transport usage versus overall travel demand. The methodology was applied to a case study in Netherlands, and was shown to be valuable in supporting tactical transit planning and decision making.

Sørensen et al. (2017) identify several technologies for measuring ridership on trains. The technologies and approaches include (1) manual counts and surveys, (2) on-board sensors, such as door passing, weight, CCTV and Wi-Fi-use, (3) ticketing systems, ticket sales or ticket validation, and (4) tracking of travellers for larger part of the journey, such as tracking of mobile phones and payments. Pelletier et al. (2011) presents an overview of the first developments of smart card. Smart cards are used to store individual data such as identification, biometrics, photos, banking data, transportation fares, etc. In transit, the main purpose of smart cards is to collect revenue, but they also produce detailed data on onboard transactions which can be useful to transit planners on both a strategic, tactical and operational level. Smart cards in public transit are usually issued by the operators to be used on their own system, and the cards are typically tapped over the reader when the user enters the vehicle.

Pelletier et al. (2011) summarise the pros and cons of smart card use in public transit which they revealed in their literature review. Some of the pros and cons are also valid for mobile phone data. Disadvantages are for instance that the data cannot provide information on trip purpose or on user assessment of service, and that development cost is high. Furthermore, the ultimate destination is not provided. Advantages include that the user role in data collection previously achieved by the survey process is minimized, as well as improved data quality and increased amount of statistics available.

Data from on-board sensors and ticketing systems are typically managed by the transportation providers. However, surveys, payments statistics and mobile phone data may be available to stakeholders outside the public transportation system, which can be an advantage because access to ridership data can be an issue for business reasons. Furthermore, mobile phone data appears to be an interesting option because they can track complete journeys.

Mobile phone data can be used to derive good estimates of dynamic quantities, such as travel times, train occupancy levels and origin-destination flows, for transportation studies (Aguilera et al., 2014). The advantages of mobile phones as sources of data include:

- the potential to generate information about travels that utilise different modes of travel (such as walking, bus, train),
- to track journeys that include transfer between trains,
- to estimate commuting patterns,
- and to derive estimates of travel times, train occupancy levels and origin-destination flows.

Several studies on mobile phone data in transportation research utilise Call Detail Records (CDR) data. This paper studies a different type of mobile phone network source that will be presented in Section 3.2.1. Three main types of mobile phone data are collected using passive collection: CDR data, Probes data and Wi-Fi data (Larijani et al., 2015). CDRs are generated by phone communication activities and contain relevant information about the activity (e.g., caller/callee, time, duration) and the location of the cell phone tower that handles the communication (Zhao et al., 2016). Studies have shown that CDR data can be used to study habits and mobility patterns of mobile users (Bianchi et al., 2016; Zhao et al., 2016), to study user movements (Leo et al., 2016), and to calculate commuting matrices with a very high level of accuracy (Frias-Martinez, et al., 2012). Studies have also looked at utilizing mobile data to estimate intra-city travel time (Kujala, Aledavood, & Saramäki, 2016) and have shown that mobile data could be employed as a real-time traffic monitoring tool (Järv et al., 2012).

Studies point out that CDR data are coarse in space and sparse in time (Becker et al., 2013) because people's phone communication activities are unevenly distributed in space and time. The bias of CDR data in human mobility research depends on what research question one wants to answer and how frequently, as well as when and where, one uses the mobile phone to contact others (Zhao et al., 2016). It has therefore been suggested that researchers should use CDR data with caution.

CDR data contain information about the caller/callee, so they are not anonymous. Consequently, studies that utilise CDR data are required to protect privacy through measures such as anonymizing the data (i.e., removing personal identification), only using the minimum of information needed for the studies, only presenting aggregated results and not focusing the analysis on individual phones (Becker et al., 2013). With pseudo-anonymised data (i.e., the ID is replaced with a code), the record must be pre-processed to reduce probability of re-identification. A common procedure is to decrease time resolution or increase space granularity (Bianchi et al., 2016). Norwegian Law states that collected personal information should only be used for the specific purpose for which it was originally collected (Drageide, 2009). As a consequence, any use of CDR data that goes beyond billing requires an active consent from the subscriber. Furthermore, the EU General Data Protection Regulation (GDPR) will be enforced in May 2018.

We have found no publications that combine data from mobile phone networks with comparable registrations of number of travellers based on on-board ridership data from train operators, even though others have combined mobile phone data with other types of public transport data such as Holleczek et la. (2014).

1.3 Research purpose

This paper studies how mobile phone data can be used to analyse the number of travellers on trains. This research has both long- and short-term perspectives. In a short-term perspective, we study how mobile phone data can be used to analyse the number of travellers on trains. In a long-term perspective, we can measure traffic flows in new ways to cover whole journeys. In addition to providing information about ridership, mobile phone data can provide information about journey flows that is not limited to each mode of transportation, but for complete journeys including travel modes such as walking, bus and train. Related to train travel, we can track train journeys that include transfer between trains, which is difficult to obtain using established techniques. Such tracking may raise personal privacy concerns, but it is not necessary to identify individual trips but to focus on flows and movements of large groups, and such data can be made anonymous (Olsson and Bull-berg, 2015). We will then be able to see transport patterns and not only measure the volume of traffic at those points where there is a count. One can also seek explanations by combining ridership data with, for example, data on punctuality or weather. We are interested in this type of data to evaluate major transport infrastructure investments such as new double tracks of railway tunnels. Several such projects are ongoing in Norway. We investigate how mobile phone data can be used in future evaluations of these projects.

The purpose of this study is to test the use of mobile phone data to measure train ridership and to investigate the potential for using mobile phone data to describe travel patterns that include train travel. Our research questions (RQs) follow.

- RQ1. Is it possible to combine mobile phone data with railway infrastructure and train traffic data?
- RQ2. What are suitable formats for presenting and analysing train ridership based on mobile phone data?
- RQ3. To what extent is the format of available mobile phone data suitable for measuring the number of mobile units passing close to the railway line?

We will discuss our results as well as methodological and technical challenges with such an approach to estimate train ridership compared to other established methods.

2 ANALYSING RIDERSHIP ON TRAINS

2.1 Use of information about number of travellers

The distribution of travel demand can be analysed and presented as a function of time or as a spatial distribution (Vuchic, 2005). Spatial distribution measures the volume of travellers in different parts of a transportation network. In a railway context, spatial distribution is used for different parts of the network. However, it would be interesting to track spatial distribution of a larger part of journeys, not just the railway ride, to ideally include the whole trip from origin to destination. The time distribution of number of travellers can be studied in different time perspectives, including variations during a day, during the week, yearly variations and long-term developments spanning several years. Daily variations in commuter transport are characterised by the morning and afternoon rush hour peaks.

Train ridership is influenced by a number of factors, including fares, transit time, transit comfort characteristics and feeder accessibility of transit, price and service characteristics of the competing modes, seasonal variations and monthly working day variations, as well as socioeconomic conditions of the service areas in the medium or long term (Doi and Allen, 1986). Demand for railway travel is typically expressed in number of travellers. Other more nuanced measures include the common format origin-destination matrices. Vuchic (2005) lists a set of relevant key performance indicators related to ridership:

- average passenger trip length, total passenger-km divided by number of passengers,
- average passenger volume, total passenger-km divided by line length,
- coefficient for flow variations to indicate the degree to which passenger volume peaks along a line,
- coefficient of passenger exchange, what proportion of passengers that are exchanged along a line,
- riding habit, how much of a population in an area that utilises the transport in question (such as commuter railway), and
- market share, use of a particular type of transportation in relationship to total travel volume in the same market.

Traditionally, there are multiple methods for calculating the demand between an origin and destination point. The most common is the O-D matrix that characterizes the transitions of a population between different geographical regions representing the origin (O) and destination (D) of a route (Frias-Martinez et al., 2012). The most commonly used method for populating these matrices is user surveys. Strengths of traditional surveys are that they include important information about the respondent, such as age and gender, and also include information about the purpose of the trip (Alexander et al., 2015). A major problem with user surveys is declining response rates (Schoeni, et al., 2013), which may introduce bias into the samples. Such surveys are typically not done with a higher frequency than yearly, but may also be conducted less frequently and not necessarily on a regular basis. Consequently, this method may possess low frequency, high cost, varying data quality, low precision and susceptibility to errors. Alternatives to traditional transport surveys include an origin-only automatic fare collection system, as proposed by Zhao et al. (2007), and mobile phone data (e.g., Jiang et al., 2013). Mobile phone data can be used to describe people's movement patterns, as illustrated in the study by Calabrese et al. (2013) who analysed the mobile phone records of a million users in Boston to describe transportation needs.

Passenger counting is the key measuring parameter associated with ridership. Different measurement types and ridership estimation techniques are applied for different network levels. The selection of the appropriate network level is dependent on the particular use and issue being addressed (Gordillo, 2006; Boyle, 1998). Gordillo (2006) and Boyle (1998) identifies uses of passenger counting and ridership calculation based on the way data are measured. The uses are different for each measurement type, and the operator normally uses multiple types to fulfil different purposes. The types and their relevant usages are as follow.

- System Level Use: Tracking system-wide ridership totals to assess changes in ridership.
- Route Level Use: Ridership by route, passenger loads at maximum load points, compiling ridership by day type and time period and monitoring schedule adherence. Performance measures are frequently calculated at the route level, and running time adjustments, route revisions, and ridership trends also rely on route-level data. Route level data are used primarily for planning and scheduling.
- Trip Level Use: Data on ridership by trip is used to add or delete trips and to adjust running times, schedule adherence and passenger loads. Link-loading between two adjacent points are important for rail capacity.
- Stop Level Use: Data on entries per station and boarding and alighting by stop are typically used in adjusting running times and in service planning in assessing route performance.
- Origin-destination level: Measures the number of passengers travelling between a pair of stations. OD data are used to help with revenue maximisation.

The Norwegian National Rail Administration (former Jernbaneverket, now BaneNor) has published annual reports on the Official Railway Statistics in Norway. The railway statistics include aggregated data on number of travellers, passenger kilometres, and number of sold single tickets and monthly tickets. The practice for measuring ridership is manual counting at chosen stations on each railway line.

2.2 Mobile phone network data

Much research has focused on developing methods to extract meaningful information about human mobility from mobile phone traces and understanding its limitations (Alexander et al., 2015). Mobile phone data can be utilised in estimating commuting patterns and travel times for individuals. Chaudhary et al. (2016) discuss collecting information about occupancy levels of public transportation system using smartphones. They show that patterns observed can predict occupancy level in a bus, with accuracy up to 92 percent. Higuchi et al. (2015) identify a number of innovative uses based on mobile devices, including several technologies that typically are found in smartphones, such as GPS, Wi-Fi, and Bluetooth. Mobile phone data sets allow for a statistical analysis of human activities at a fine level of detail (Leo et al., 2016).

Various approaches can be utilised for calculating this information by analysing the exchange of information between the mobile base station and cellular network. Most studies perform some kind of trip extraction to extract the movements relevant for traffic analysis from the raw cellular network data (e.g., Calabrese et al., 2011; Doyle et al., 2011; Alexander et al., 2015; Iqbal et al., 2014). Because cellular network data can contain a lot of noise, there is no obvious definition of what a movement/trip is (Gundlegård et al., 2016). Hence, trip extraction algorithms vary a lot among different authors.

An origin-destination matrix can be computed based on the extracted trips (e.g., Calabrese et al., 2011; Larijani et al., 2015). For instance, Alexander et al.'s (2015) method estimates average daily origin-destination trips from triangulated mobile phone records of millions of anonymized users. The CDR records are converted into cluster locations and inferred to be home, work or other depending on observation frequency, day of week and time of day. The aggregation of OD flows gives an estimate of the number of cell phone users who are travelling, but only those of the operator that provided the data (Gundlegård et al., 2016). As a result, this can only give information about how the travel demand distributes relatively between different OD pairs. To estimate the total travel demand in terms of the number of people travelling, authors use different scaling factors (e.g., Alexander et al., 2015; Iqbal et al., 2014; Toole et al., 2015; Calabrese et al., 2011).

Several authors have also tried to reconstruct the specific travel mode and route that a user took for a trip, which is challenging. However, as Larijani et al. (2015) showed, detection of the trip segments in which people take the metro is promising, because underground tunnels are served by dedicated base stations (Gundlegård et al., 2016). Xu et al. (2016) used a large-scale mobile phone data set to estimate demand of bicycle trips in a city. Another approach is to use smartphone travel surveys based on smartphone applications to capture accurate details about individuals' travel behaviour, as presented by Assemi et al. (2016). However, extracting required information (e.g., travel mode and purpose) from the data captured by smartphone applications is relatively complex. Holleczek et al. (2014) showed that urban mobility patterns and transport mode choices can be derived from mobile phone CDR coupled with public transport data. This public transport dataset consists of trips made by 4.4 million anonymized users of Singapore's public transport system. The advantage is that passengers in Singapore use smart cards when getting on and off trains and buses, hence the data include station and time of departure and arrival of each trip.

Calabrese et al. (2013) show that mobile phone data can be used to describe people's movement patterns, as an alternative to traditional transport surveys. They obtain mobile phone records of a million users in Boston during a three-month period to describe transportation needs. They discuss three challenges using mobile data. The first is that demographic information about individuals was not available due to privacy concerns. Secondly, mobile users were not necessarily representative of the whole population. Thirdly, the data were not formatted for this type of analysis. To address the first challenge, they used aggregated data in which users were collected into groups corresponding to the most detailed level of economic and demographic data that were available. The second issue introduces sample bias amongst the population. To validate the representativeness, Calabrese et al. (2013) calibrated the data based on information from security inspections of the vehicles, which included mileage condition, to check if the estimated mileages seemed realistic. Mobile phones are a growing data source through activated apps. Apps can track complete journeys, especially if the users have allowed apps to use GPS for tracking. Such apps can be supplied by private or public transportation entities, or they can be apps for navigation, health monitoring or other types. Data from these apps are typically managed by the organisation issuing the app, and not by the mobile network managers.

There are both advantages and disadvantages by using mobile phone data. For instance, CDR data contain approximate locations when the phone communicates with a cell phone tower, hence providing an inexact and incomplete picture of daily trips. Furthermore, the mobile phone data are not able to provide information about the traveller, like age, income or purpose of trip, as a survey would (Alexander et al., 2015). On the other hand, mobile phone data are automatically collected, which makes them more frequent and economical than, for instance, a survey. In addition, as mobile phone data can be gathered over a longer time period, it can capture information such as variations in the travellers' daily travel behaviour (Alexander et al., 2015).

Having established the need and uses of ridership data, the short-term vision of this project is to investigate if available mobile data and railway traffic data are a viable method for calculating number of travellers on a specific train route. This data can serve as a source of validation, quality assurance and triangulation for the currently available data in railway industry.

3 APPROACH

Three sets of data were obtained for analysis purposes: punctuality data for specific train stations, mobile phone data in a specific format from adjacent cell sites, and actual passenger counts from the trains. The data sets are described in Section 3.2. The purpose of this study was to investigate suitability of mobile data to find ridership on trains and to investigate if it is possible to combine mobile data with railway infrastructure and train traffic data.

The first part of the analysis focused on combining mobile telephone data with data describing the railway and railway traffic. As we will see, a key issue is to relate peaks in mobile telephone connections to the passing of trains. The final part of the analysis utilises counts of passengers on the trains and connects the number of travellers on the trains to the number of mobile phone connections.

The analysis was done in the following steps:

- identifying base stations,
- connecting trains to base stations,
- graphic inspection of handset counts in relationship to trains passing the base station,
- analysis of data resolution, comparing data sets of five- and one-minute collection time intervals,
- statistical analysis using a proposed algorithm,
- extracting the peaks in handset count, correlated with trains passing the base station, and
- comparison and validation with actual ridership data.

A key step of the approach was to locate suitable mobile phone base stations close to the studied railway line. Thus, the approach of selecting base stations is described subsequently, along with analysis methods. To begin with, the following section describes the use case.

3.1 Use case description

The railway line selected as case for this preliminary study goes into a city in Norway, where people commute daily to work from towns on the outskirts of the city. The analyses look at five base stations located near the railway tracks and in connection with five of the train stations on the selected railway line. To anonymize the data, we denote the train stations as U, V, W, X and Y, where station U is farthest away and station Y is closest to the city, as illustrated in Figure 1. The base stations are denoted B1, B2, B3, B4 and B5.

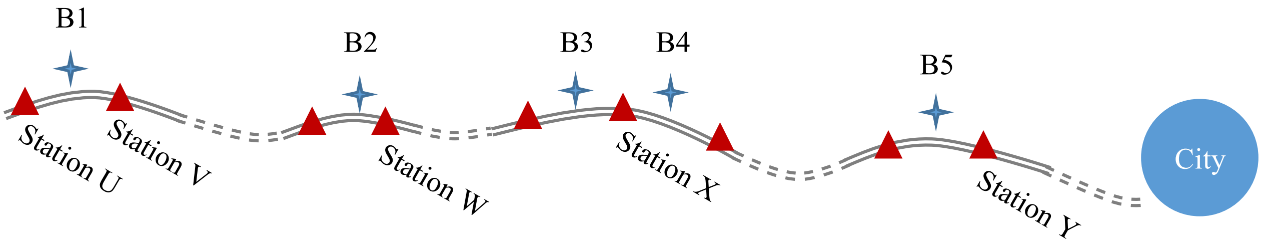

Figure 1. Illustration of the studied railway line, with the five base stations in relationship to the train stations.

Data were collected in three time periods:

1. eight consecutive days in the spring of 2016,
2. three consecutive days in the fall of 2016, and
3. nine consecutive days in the spring of 2017.

Punctuality data were made available for all three time periods. The mobile phone data in time period 1 were collected with five-minute time intervals between each collection time. In time period 2, the mobile data were collected with one-minute time intervals between each collection time. This data set is less complete than the data set with five-minute collection intervals. In time period 3, the mobile data were collected with one-minute time intervals between each collection time. This data set is complete. Automatic Passenger Count (APC) data were made available for time periods 1 and 3.

3.2 Available data and research material

3.2.1 Mobile phone data

The mobile phone data sets are counts of the number of handsets recorded at the selected base stations. Since the base station has limited range and a mobile subscriber can be connected to one of six base stations based on the signal strength, due to a hexagonal symmetry for frequency reuse (Mac Donald, 1979), the handset counts for a certain base station serve as an indicator of the number of people using the mobile network in the coverage area of the base station.

The counts are the total number of connections to a base station cell, and the data set consists of *collect_time*, *cell_id*, and *count_handsets*, meaning the number of handsets connected to a base station cell at a given point in time. Example of data format is shown in Table 1. For the purpose of this study, a script was made that extracts these data with certain intervals. When a phone turns up on another base station, it will no longer be counted on the previous one. The handset count will serve as an indicator of the number of people using the mobile network in the coverage area of the base station, at the exact time the data are collected.

It is worth mentioning, though obvious, that passengers without mobile phones and those whose mobile phones are switched off or not working, will not be recorded by this approach. In the source system used to extract the count data, it is not possible to see each individual event. However, at any time, the mobile network operator can extract the number of mobile phones that were last seen on the base station.

The mobile phone data used in this study are pure counts by cell by time unit, which are anonymous data that are not covered by the privacy legislation and cannot be used to identify a person. Both from an ethical and legal point of view, it is important to protect personal information and respect people's privacy. Data that do not include personal information are basically unproblematic, both as individual data sources and the combination of several sources. Combination of different data sources in which persons are the link between the various data is more problematic. Data from different sources can be combined with personal data without revealing personal information, but this can be challenging. Anonymity in datasets is typically achieved by aggregation, in which each group includes so many persons that individuals cannot be identified. To make sure that we do not violate private privacy, the data used in the study have been approved by privacy representatives.

Table 1. Sample table showing the data format of mobile phone data.

Collection time	Cell id	Cell easting	Cell northing	Handset count
2017-06-01 06:19:00	xxx	xxxx	xxxx	159

3.2.2 Railway traffic and infrastructure data

The Norwegian National Rail Administration (Jernbaneverket at the time of the study, now Bane NOR) records punctuality data for individual trains at arrival and departure at stations. Data are recorded in a database (TIOS), of which we have obtained a copy. The data describe the movements of the trains through the network. These records include scheduled and actual arrival and departure times for each train at every station, train number and operating company, and class information (e.g., freight, running empty, or passenger train). Example of data format is shown in Table 2.

The physical layout of the Norwegian railway network is described in several formats, many of them being available on the internet (BaneNor, 2017). We have utilised this information to combine train data and mobile phone data. In particular, we calculated when trains passed close to the mobile phone base stations.

Table 2. Sample table showing the data format of TIOS data.

Date	Train no.	Station code	Scheduled arrival	Actual arrival	Scheduled departure	Actual departure
2017-06-01	815	XYZ	17-06-01 11:59	17-06-01 11:59	17-06-01 12:00	17-06-01 12:00

3.2.3 Automatic Passenger Count (APC) data

The third data set is passenger counts from trains based on Automatic Passenger Counting (APC). The APC system is installed on a sample of the vehicles on the railway line studied in this work. The APC data that were made available were collected from the trains on two of the railway lines that pass station Y, which we denote Line 1 and Line 2. The APC system registers the number of people who board and alight through each train door on every station by means of sensors in the doorways. Norsk Regnesentral (Norwegian Computing Centre) has developed a mathematical tool that, based on the APC data, uses a statistical model to calculate the total number of passengers on each train at different points in time (Teknisk Ukeblad, 2014). The data set that was made available to us is the calculated total number of passengers on each train when the train is leaving the station. Example of data format is shown in Table 3.

Table 3. Sample table showing the data format of APC data.

Train no.	Product name (start-stop)	Direction	Date	Day	Location (train station)	Sum boarding	Sum alighting	Sum passengers
815	Small town-Other town	Towards Other town	17-06-01	Thu	City C	37	32	189

3.3 Approach of selecting base stations

The base stations in close proximity to the railway tracks were located based on coordinates and description of coverage. The Quantum geographic information system (QGIS) was used to import the coordinates for the base station, together with a map of the area and the railway lines. The initial criteria used to select base stations were that the base station should be within 2 km from the tracks, but also excluding base stations more than 1 km from the end train station in the opposite direction. Generally, we also excluded cells with descriptions of indoor coverage. These restrictions resulted in approximately 600 cells divided between around 100 base stations. To arrive at a more appropriate number of cells for our preliminary analysis, we singled out the base stations that with greater certainty would be connected to train travellers, e.g., including base stations with descriptions containing railway, train stations or railway tunnel. The number was further narrowed down to about 10 by selecting base stations that (1) were located between train stations, (2) were not located near a main road, and (3) were located in more deserted areas. Mobile phone data were obtained from the selected base stations and compared with timetable data. The selected base stations are located so that it is likely that a large share of the mobile phone traffic is generated by train passengers.

3.4 Applied methods

As mentioned in Section 3.3, the selected base stations were located between the train stations on the case line. A first step in the analysis was to compare the handset counts on the base stations with when trains are passing the location in proximity to the base station to see if it is possible to connect passing of trains with possible jumps in the handset count data. The train traffic data give the actual time of when the trains arrive and depart from the train stations. We therefore needed to find how long after the train leaves the train station the train passes the location in proximity to the base station.

We looked into two different approaches to find the approximate time of when the train passes the base station. The first approach is based on the distances and times between the stations, with the assumption of constant speed. The second approach is based on the allowed speed on the railway lines.

Approach 1: With the assumption that the train has an average constant speed from station A to station B (see Figure 2), the time s can be expressed as $s = \frac{y}{x}t$ minutes. The travel time from station A to station B can either be the scheduled travel time or the actual travel time for a specific train at a specific time. We used actual travel time when it was available in the train traffic data.

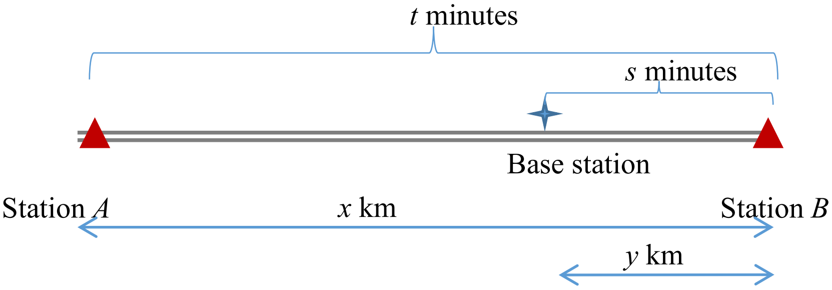

Figure 2. Illustration of a base station located between two train stations and the notations used to calculate travel time of the train from train station B to the base station.

Approach 2: Based on the speed limits on each section of the line, which are rarely constant between two stations, the average speed v from the nearest train station to the base station can be calculated, and the time is found to be $s = \frac{y}{v}60$ minutes.

Based on approach 1 or approach 2, we could then estimate the time for when trains passed the base stations, utilizing data that show actual train movements. If the train is traveling from A to B, the point in time when the train passes the base station is then given by $t_{arrival}(B) - s$. If the train is traveling from station B to station A, the point in time when the train passes the base station is given by $t_{departure}(B) + s$. Based on preliminary analyses, the approaches yielded quite similar results. We therefore chose to use approach 1 throughout this study.

3.4.1 Algorithm

This section presents an algorithm that compares the collection times with the calculated approximate times the trains pass the base stations. The method aims at quantifying the impact of a train passing and its extent. The main point of interest is to categorize the collection times for mobile data to reflect whether a train has passed the base station or not. This basic categorization distinguishes the closest count collection points to the passing of a train and their adjacent values, with the objective of determining if a train has passed between two subsequent collection times. Further sub-categorization is made to analyse the effects of rush hours and direction.

The algorithm, as presented below, compares each single collection time with an array containing the calculated times of all trains that pass the base station. The collection time closest to the time of train passing is the minimum value of the time difference between the collection time and the train time. When the value is less than the collection time interval, that implies that the specific collection point is adjacent to a train passing. Otherwise, if the value is greater than the collection time interval, the specific collection time instant is not reflective of a train passing.

The input values of the algorithm are the three known variables: number of handsets (numerical value), denoted *count_handsets*, collection times (date and time value), denoted *collect_time,* and time of train passing at base station (date and time value), denoted *train_time*. In addition, the collection time interval is denoted t_{CI}. The output variables calculated in the algorithm are illustrated in Figure 3:

- $D(i)$: time difference between collection time i and the nearest train passing time before it (value in minutes);
- $T(i)$: string value "Yes" or "No", indicating if a collection time is adjacent to a train passing since the last collection time. 'Yes' if a train has passed within the collection time interval (i.e., $D(i)<t_{CI}$), 'No' otherwise, i.e., $D(i)>t_{CI}$; and
- $I(i)$: percentage increase in handset count, i.e., percentage increase from the previous count at a specific collection time, given by $I = \frac{d_c^i}{count(i-1)}100$, where d_c^i is the difference in handset count given by $d_c^i = c_i - c_{i-1}$.

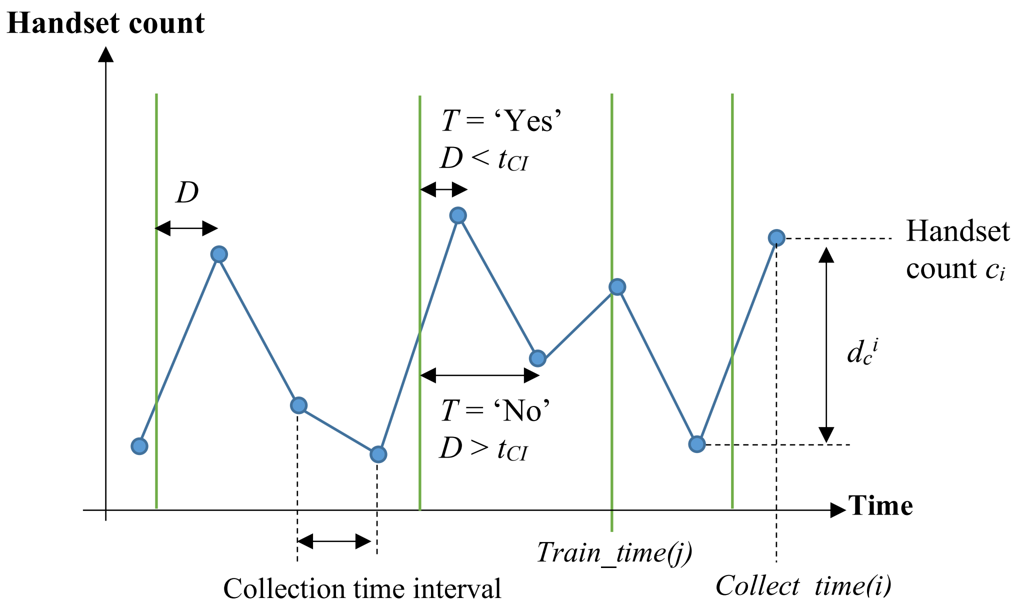

Figure 3. Illustration of the output variables of the algorithm. The variable D is the difference in time between a collection time and the nearest train time. T is the string 'Yes' or 'No', whether D is less than the collection time interval or not. Handset count c_i at collect_time(i) and difference in handset count d_c^i.

The steps of the algorithm are as follows.

1. Find the length of the arrays *collect_time* and *train_time,*

   n = length(*collect_time*); m = length(*train_time*).

2. **For** i in 1 to n; **for** j in 1 to m: compare the value *collect_time*(i) to the value *train_time*(j). If the value *collect_time*(i) is greater than *train_time*(j), subtract it from the collection time value and calculate the difference in minutes. The value is saved in a temporary array P. That is,

   If *collect_time* (i) > *train_time* (j)
   $P(j)$ = *collect_time* (i) - *train_time* (j).

```
3. Calculate the closest train passing in minutes, i.e., time difference $D(i)$, for a specific
   collection time $i$ as the minimum value in $P$.
         $D(i) = \min(P)$
4. Step 2 and 3 are repeated for each $i$, which will result in a difference value array $D$ of length
   $n$.
5. For each $collect\_time(i)$,
         If the corresponding Difference value is less than $t_{CI}$ minutes: $D(i) < t_{CI}$
               then the $T$ value is Yes: $T(i)$ = 'Yes'
         else if it is greater
               $T(i)$ = 'No'.
6. For $i$ in 1 to $n$, the percentage increase $I(i)$ in each handset count is calculated from the
   previous count value as,
         $I(i) = \frac{c_i - c_{i-1}}{c_{i-1}} 100.$
```

The described algorithm was implemented in MATLAB R2017b (9.3), a numerical software environment and programming language developed by MathWorks. The values calculated in the algorithm were then analysed in the statistical software R 3.4.

3.4.2 Graphic inspection

The graphic inspection entails plotting the handset counts against time. The handset counts were compared to the calculated approximate time of when trains are passing the base station, illustrated by vertical lines in Figure 4. The assumption is that as the train pass the base station, the number of users connected to the specific base station increases, corresponding to the people travelling on the trains, because the passengers on the trains also get connected to the specific base station and thereby increase the count at the next measurement point, so that at the next collection time, a higher value is observed.

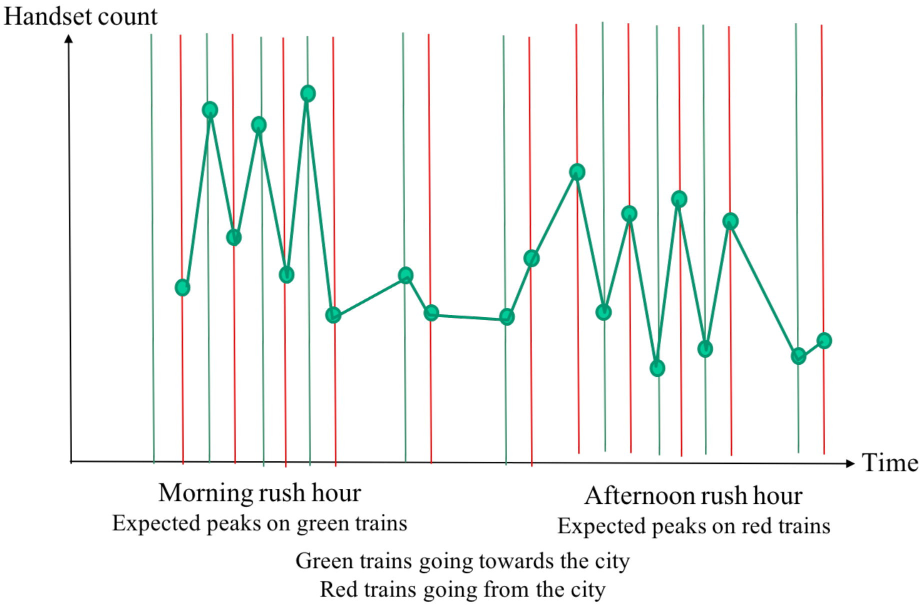

Figure 4. Illustration of expected correlation between the handset counts and the morning and afternoon rush hours, where the vertical lines represent the calculated times the trains are

passing the base station. Trains travelling towards the city are illustrated by green vertical lines, and trains travelling away from the city are illustrated by red lines.

The directions the trains are travelling in are included in the analysis, where the trains are divided into two categories: going towards the city, and going away from the city. The direction of the train is illustrated in Figure 4 by green lines for trains travelling towards the city and red lines for trains travelling away from the city. Morning rush hour trains going into an urban centre were expected to have more passengers than the trains going in the other direction, especially in typical suburban and rural areas. We would therefore expect peaks in the diagrams when trains going into the city pass a mobile base station, as illustrated in Figure 4. The opposite would be expected in afternoon rush hour. We would expect a lower variation in periods when there are no trains, compared to when trains pass.

3.4.3 Collection time interval

The collection time interval of how frequently the handset counts are collected is expected to have an effect on the results of the analysis. The collection time interval was studied by aggregating a data set with one-minute collection time intervals down to a data set with five-minute intervals. Because the handset counts are snapshots of the number of handsets connected to the base station, the aggregation is done by extracting every fifth measure collected at the times 00, 05, 10, etc., past each hour to create a new data set with five-minute collection time intervals. These two data sets were compared by graphic inspection and compared to the calculated times the trains passed the base station.

3.4.4 Statistical analysis—Violin plot

The statistical analysis was based on the categories presented in Table 4 and Table 5.

Table 4. General categories for statistical analysis and the values included in the analysis.

Category	Condition	Values	Description
Yes	$D(i) < t_{CI}$ $T(i)$ = 'Yes'	*collect_time*(*i*), *count_handset*(*i*), *I*(*i*)	All collection points and adjacent handset counts, as well as the percentage increase in handset count, where a train has passed since the previous collection time.
No	$D(i) > t_{CI}$ $T(i)$ = 'No'	*collect_time*(*i*), *count_handset*(*i*), *I*(*i*)	All collection points and adjacent handset counts, as well as the percentage increase in handset count, where a train has not passed since the last collection time.

Table 5. Categories for statistical analysis within the category of events where a train has passed since the previous collect time, i.e., $D(i) < 1$, $T(i)$ = 'Yes'. Direction towards the city is denoted by true (TR) and false(FL).

Category	Conditions		Description
	Collection time (hrs)	Direction	

TR	All day	Towards the city	All trains travelling towards the city
FL	All day	Away from the city	All train travelling away from the city
MRTR	06:00 to 09:59	Towards the city	Trains travelling towards the city in the morning rush hours
MRFL	06:00 to 09:59	Away from the city	Trains travelling away from the city in the morning rush hours
ERTR	15:00 to 18:59	Towards the city	Trains travelling towards the city in the evening rush hours
ERFL	15:00 to 18:59	Away from the city	Trains travelling away from the city in the Evening rush hours

To validate and compare the results of the statistical analysis, certain assumptions need to be considered. The assumptions are based upon the general trends believed to exist based on previous general traffic patterns.

1. The train passing causes an increase in the number of handsets in addition to the natural variation of the mobile handset count, i.e., the 'Yes' category is higher than the 'No' category.
2. In the morning (MR) and evening rush hours (ER), the number of travellers on the train is higher in count than at other hours of the day. This means the 'MRTR', 'MRFL', 'ERTR', and 'ERFL' categories possess higher statistical values than 'TR' and 'FL'.
3. In the morning rush hour, the number of passengers travelling on trains towards the city is higher than the number of passengers on trains travelling away from the city, i.e., 'MRTR' statistical values are higher than 'MRFL'.
4. In the evening rush hour, more passengers travel away from the city than travel towards the city, i.e., the 'ERFL' category has higher value than 'ERTR'.

Information about the distribution of the counts in each of the categories can be visualised with violin plots. The violin plot is a combination of boxplot and a kernel density plot to reveal structures within the data (Hintze & Nelson, 1998). The box plot shows four main features of a variable, centre, spread, asymmetry and outlier, which are also included in the violin plot. Similar to box plots, violin plots allow us to compare and visualize the relationship between numerical and categorical variables. In addition, the violin plots have a rotated density distribution on each side, showing the distributional characteristics of batches of data.

3.4.5 Extracting the peaks in handset count

Apart from the peaks, we expect that the handset count data will show a varying trend throughout a day and a week. We investigated therefore methods to extract the peaks. This would remove the variation contributed, for instance, from the people within the range of the base station cell who are not on the train. Two approaches were tested. The first extracts the daily variations to find the extent of the peaks. This was done with a simple moving average (SMA) in R. This method will not give the exact value of the peaks, so a more accurate method will be preferred in later analyses, but for this preliminary study, we show results of the SMA analysis. The second approach to extract the peaks is to calculate the difference in handset count for each collection time as the difference between the current and the previous handset count value, that is, $d_c^i = c_i - c_{i-1}$.

3.4.6 Comparison to actual ridership

The number of travellers from the APC data was compared to the handset counts by calculating the ratio for each passing train. The algorithm was used to connect the collection times with the trains. The ratio was calculated as handset count divided by APC count. The results of this comparison are shown in Section 4.5.

4 RESULTS

The following subsections present the results from each analysis in the study. The number of handsets and the timings of the trains passing adjacent to the base station were analysed systematically. Having calculated the approximate times for trains passing at the base station, the trends are plotted, and general analysis based on the correlation of trains passing and counts variation is undertaken. The trends at different times of the day are studied separately. Moreover, the train direction is added, and analysis for trains travelling in different directions is considered. In the statistical analysis, categorization of data is based on the passing of trains and calculation of different statistical values to compare and contrast the results. The availability of the actual passenger count from the train operator served as a mechanism for validation of the ratio between the number of passengers and number of handsets on the adjacent cell sites. The handset counts are indexed in the presented analyses to anonymize the data.

The analysis is divided into a number of steps:

1. graphic inspection of peaks in handset counts in relationship to trains passing the base station,
2. analysis of data resolution, comparing data sets of five- and one-minute collection time intervals,
3. statistical analysis of the output values of the proposed algorithm,
4. extracting the peaks in handset count, analysing the extracted peaks using the proposed algorithm, correlated with trains passing the base station, and
5. comparison and validation with APC data.

4.1 Graphic inspection

Figure 5 illustrates the actual handset counts with five-minute collection time intervals plotted against time on base station B1, located between station U and station V. The time period is one day from six in the morning to midnight. The handset counts are compared to the approximate time when trains are passing the base station, represented by the vertical lines. The handset count data have some distinct peaks that seem to coincide with some of the trains passing the base station.

Figure 6 shows an example of morning rush hours on a Thursday and afternoon rush hours on a Friday on base station B1 between station U and station V. In this example, the expected pattern is not visible (see assumption 3 and 4 in Section 3.4.4). For this base station, the peaks mostly coincide with the trains traveling towards the city, regardless of time of the day.

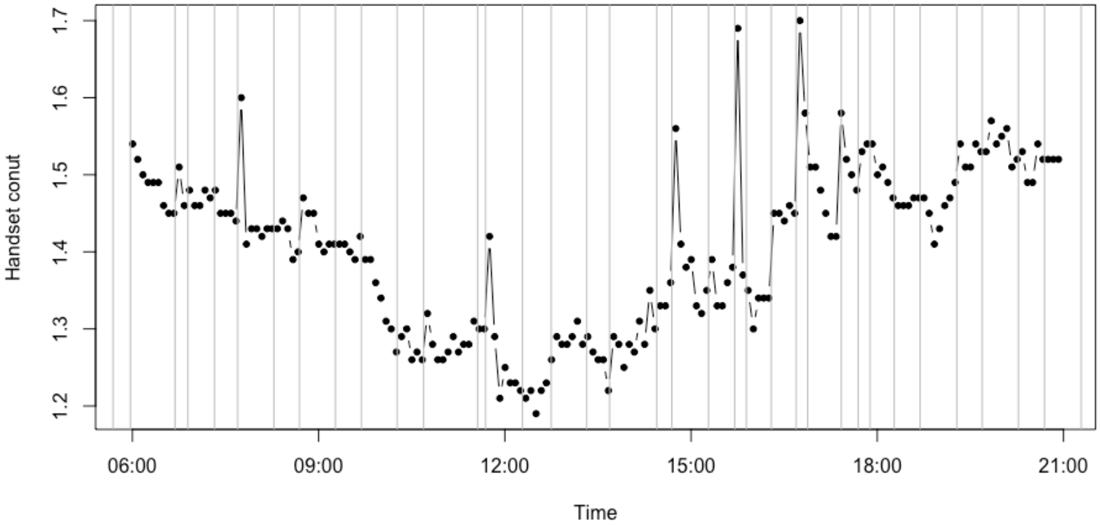

Figure 5. Handset count collected on a Friday on base station B1 located between station U and station V, from the first data set with five-minute collection time intervals. The vertical lines represent the calculated times when trains passed the base station.

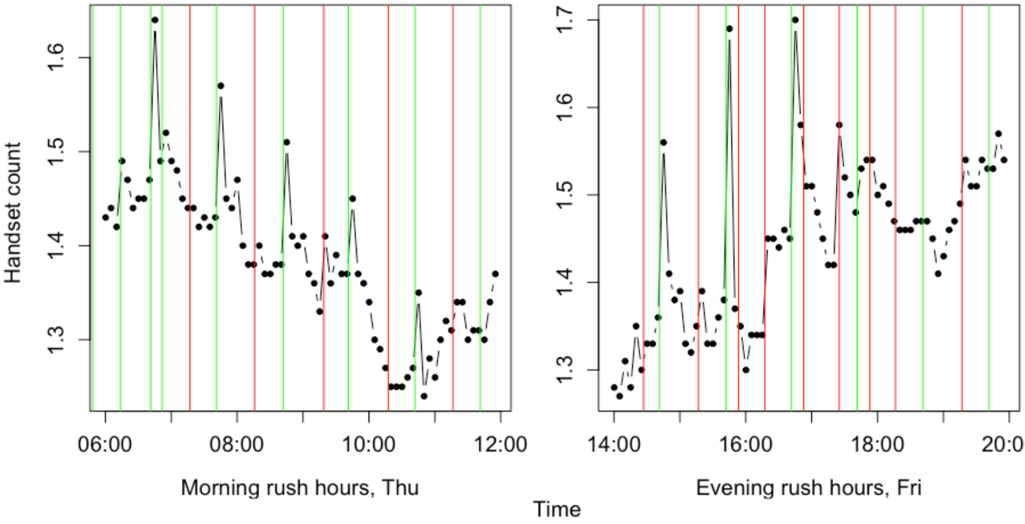

Figure 6. Handset count from the morning and afternoon rush hours on base station B1 located between station U and station V, from the first data set with five-minute collection time intervals.

4.2 Collection time interval—five minutes versus one minute

In Figure 7, the handset counts with one-minute collection time intervals is compared to the aggregated data set with five-minute intervals. Figure 7 (a) shows the data set of one-minute collection time intervals on base station B5 before station Y, Cell 2, from the second data set. Figure 7 (b) shows the aggregated data of the handset count with five-minute intervals for the same data set on the same day and base station. The handset counts with one-minute collection time intervals in Figure 7 (a) are closely related to the passing of trains. The five-minute aggregated handset counts in Figure 7 (b) show fewer and less distinct peaks, with lower values, and are less obviously relatable to the trains passing the base station. With one-minute collection time intervals as shown in Figure 7 (a), the peaks are more frequent and show that a large share of the peaks is in connection with trains travelling away from the city in the afternoon.

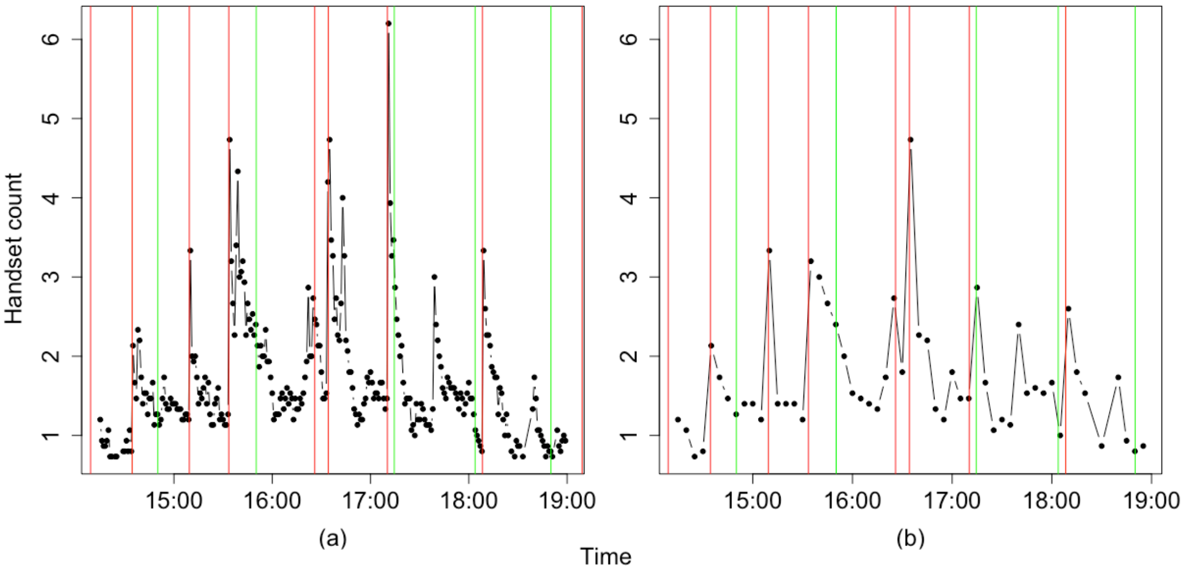

Figure 7. Handset count with one-minute collection time intervals in (a) and the one-minute interval handset count aggregated to five-minute intervals in (b) on base station B5 before station Y, Cell 2, from the second data set. Times when trains are passing are represented by the vertical green lines (towards the city) and red lines (from the city) for Line 1.

4.3 Statistical analysis

This section analyses the output of the algorithm presented in Section 3.4.1. The analysis is done on base station B5 that is located before train station Y, which is the station with the most frequent trains passing on the studied railway section. The one-minute collection time interval data set is used for a time period of five days, i.e., Monday-Friday. The output is studied based on the categories described in Table 4 and Table 5, which are direction (trains travelling towards or away from the city), time of day (morning or evening rush hour), and whether or not a train passed between the collection times (category Yes or No). The assumption is that the count values will be higher in the collection time intervals when trains are passing the base station (assumption 1 in Section 3.4.4).

Figure 8 shows violin plots of the categories described in Table 5. The density distribution of handset counts over five days is given on the y-axis. The time of day on the x-axis shows the categories of morning rush hour, evening rush hour, and the rest of the day. Which direction the trains are travelling in, i.e., towards the city or away from the city, is given by the colours as described in the legend. These are handset counts categorised in the 'Yes' category, identified by the proposed algorithm as collection times when a train has passed the base station since the previous collection time. In addition, the 'No' category when no trains had passed the base station since the previous collection time is included and categorised by the time of day. The median values in handset count for each of the categories are shown by the white circle. The black lines give the standard box plot for each of the categories. Violin plots with higher or lower values than the box plot whiskers, for instance, as the maximum value in the evening rush hour towards the city, indicate outlier(s).

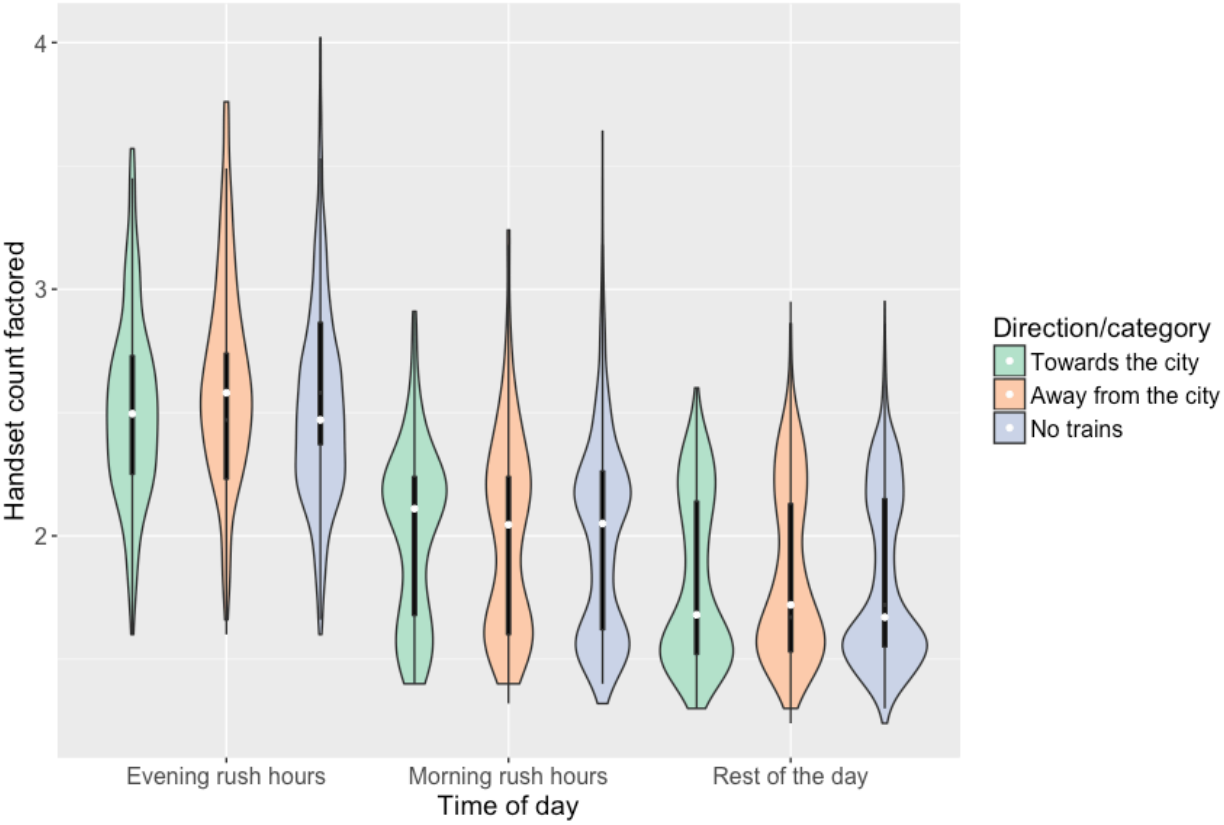

Figure 8. Violin plots showing the density distribution of handset counts over five days (y-axis), with median value (white circle) of the categories of (see Table 5) direction (x-axis) and rush hour (colour in legend) when trains had passed within the collection time interval prior to the collection time and time of day for category 'No' when no trains had passed since the previous collection time (see Table 4). The handset counts, which are represented in the figure as factors, are from the third data set collected at base station B5, with one-minute collection time intervals.

The evening rush hour in Figure 8 is the category with the highest median handset counts, compared to the other hours of the day. The direction away from the city in the evening rush hour has a median that is slightly higher than the other subcategories. In the morning rush hour, it is the direction towards the city that has the highest median, but just barely. The density distributions for the evening rush hour show one local maxima for each of the three categories, in which the category with no trains contains the highest collected handset count. On the other hand, the 'No' category shows a distribution with higher density at the lower part of the violin plot. Both the morning rush hour and the rest of the day show bimodal distributions, with two local maxima in each of the density functions. This means that within these categories there are high densities of both low handset counts and high handset counts. The morning rush hour towards the city shows a distribution with higher density at the upper part of the violin plot, i.e., the largest portion of the handset counts in this category has high values. All three categories in the rest of the day have distributions with the highest densities at the lower parts of the violin plots.

Summarising the handset counts over five days, Figure 9 shows that the average peaks in handset count is higher in the evening than in the morning. It also appears that the peaks are more frequent in the evening rush hour. Figure 9 also shows that the average daily variation is quite distinct between rush hours and the middle of the day.

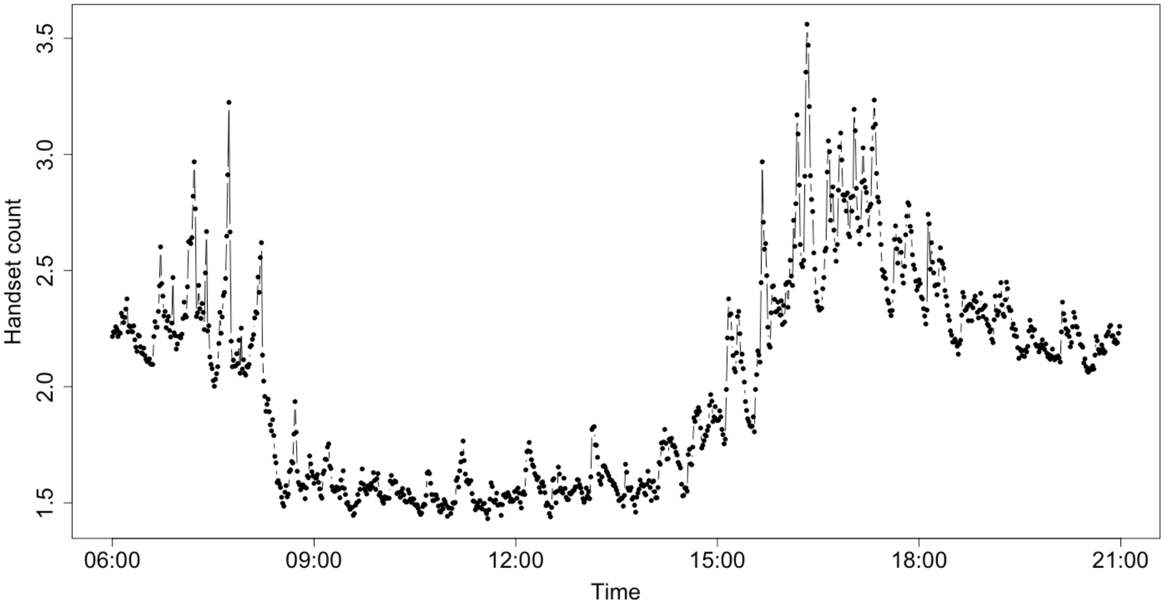

Figure 9. The sum of handset counts over five days (Monday-Friday) on base station B5 located before station Y from the third data set with one-minute collection time intervals.

4.4 Extracting the peaks in handset count

The handset counts fluctuate throughout a day, with smaller fluctuations on the weekends, as shown in Figure 9 and the count part of Figure 10. Since there is a variation through the day, we investigated methods to extract the peaks, presented in Section 3.4.5. The approaches are analysed with the handset counts collected on base station B1 for the first time period of collecting data, with five-minute collection time intervals. The result of the simple moving average is given in Figure 10, showing the handset counts, the trend line and the irregular values, which are simply handset counts minus trend values. The irregular values show jumps in the counts, presumably when trains are passing the base station.

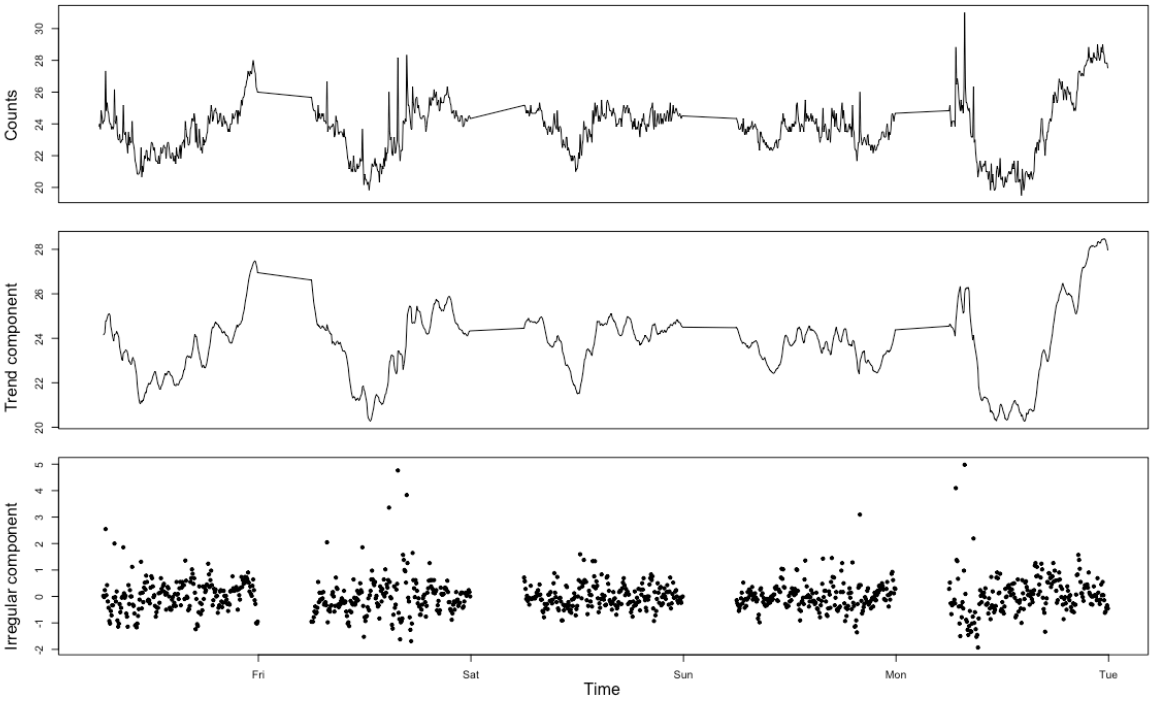

Figure 10. Counts, trend component and irregular component at base station B1 between station U and station V for six days in 2016, with five minutes as collection time intervals.

The second method of extracting the peaks as the difference in handset count, $d_c^i = c_i - c_{i-1}$, is shown in Figure 11. The difference in count, d_c^i, is the same value used to calculate the percentage increase in handset count *I* in the algorithm in Section 3.4.1. The percentage increase is analysed with violin plots in Figure 12 for data collected at base station B5 from the third data set with one-minute collection time intervals for a time period of five days, i.e., Monday-Friday. The violin plots in Figure 12 show the density distribution of the percentage increase on the y-axis for each of the categories described in Table 5. The time of day on the x-axis shows the categories of morning rush hour, evening rush hour, and the rest of the day. Which direction the trains are travelling in, i.e., towards the city or away from the city, is given by the colours as described in the legend. These are percentage increases categorised in the 'Yes' category, identified by the proposed algorithm as collection times when a train has passed the base station since the previous collection time. In addition, the 'No' category when no trains have passed the base station since the previous collection time is included, categorized by the time of day. The median values in percentage increase for each of the categories are shown by the white circles. The black lines give the standard box plot for each of the categories.

The percentage increases categorised as direction away from the city in Figure 12 have highest median values for all times of the day. The other categories have medians that are negative or close to zero. Most of the violin plots have a density distribution around zero with one local maxima, except the evening rush hour towards the city, which has a large portion of negative values. However, this category also has the highest registered percentage increase. The category when no trains passed shows density distributions around zero with both positive and negative values and a few high percentage increases.

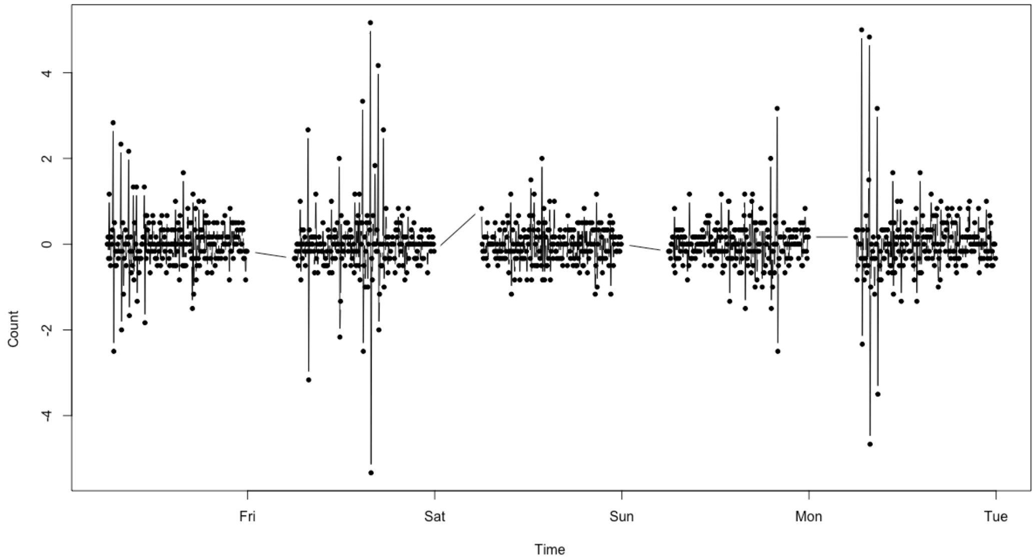

Figure 11. Difference in count, $d_c^i = c_i - c_{i-1}$, for the basestation located between station U and station V for six days in 2016, with five minutes as collection time intervals.

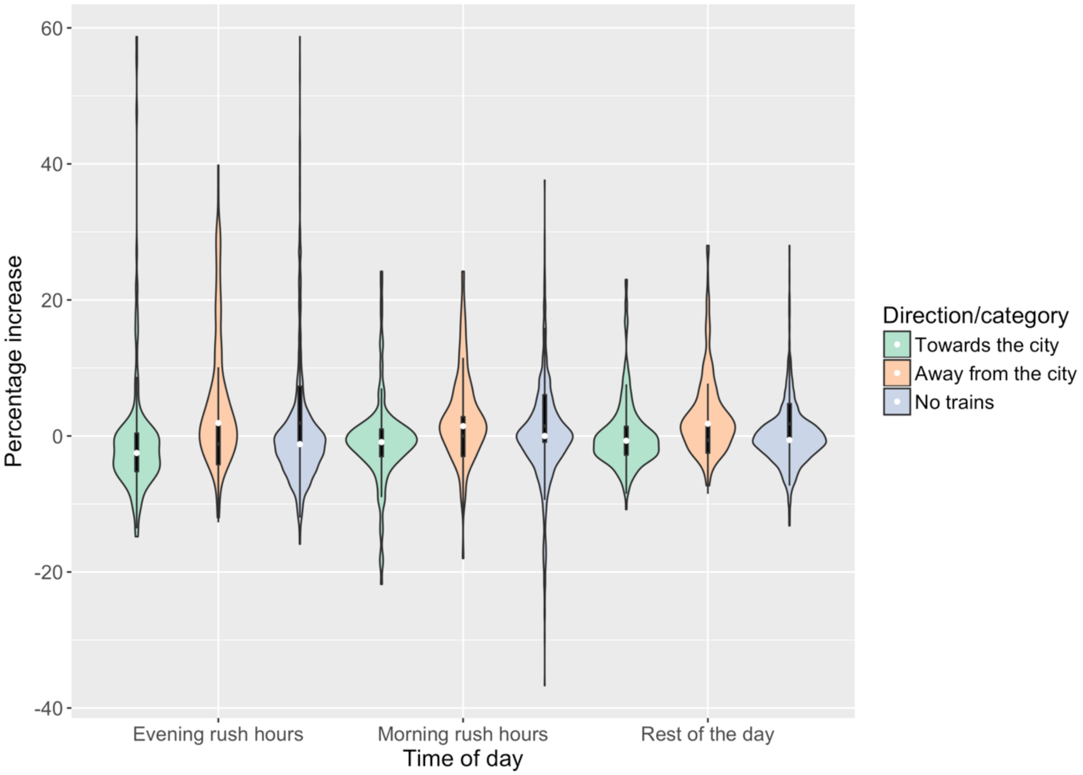

Figure 12. Violin plots showing the density distribution of percentage increase I over five days (y-axis), with median (white dot) of the categories (see Table 5) direction (x-axis) and time of day (colours in legend) when trains had passed within the collection time interval prior to the collection time and time of day for category 'No' (see Table 4). The percentage increases I in

handset counts are from the third data set, collected at base station B5 with one-minute collection time intervals.

4.5 Comparison to actual ridership

This section presents the comparison and validation of the handset counts with APC data. The passenger counts from the APC data are plotted in Figure 13 for two of the train lines as the trains passes base station B5, towards the city of Line 1 shown by the green line and away from the city of Line 1 and Line 2 shown by the red lines. The y-axis gives the APC count, however the values are indexed to anonymize the data. Figure 13 shows a clear daily variation in which most passengers are travelling towards the city in the morning and away from the city in the evening, with the exception of a few trains away from the city in the morning on Monday and Thursday, and a few trains towards the city on Thursday evening. Figure 14 show the handset counts on base station B5 from the third data set with one-minute collection time intervals. Line 1 and Line 2 are shown in Figure 14 by the vertical lines of same colours as in Figure 13. The handset counts show a daily variation in Figure 14 similar to Figure 9 with distinct variation between the morning, the middle of the day and the evening. Figure 14 also show that the frequency of peaks in handset counts are higher in the evening rush hours compared to the morning rush hours. This result matches the observation from the violin plot of the percentage increase, in which there were most positive increases for trains travelling away from the city.

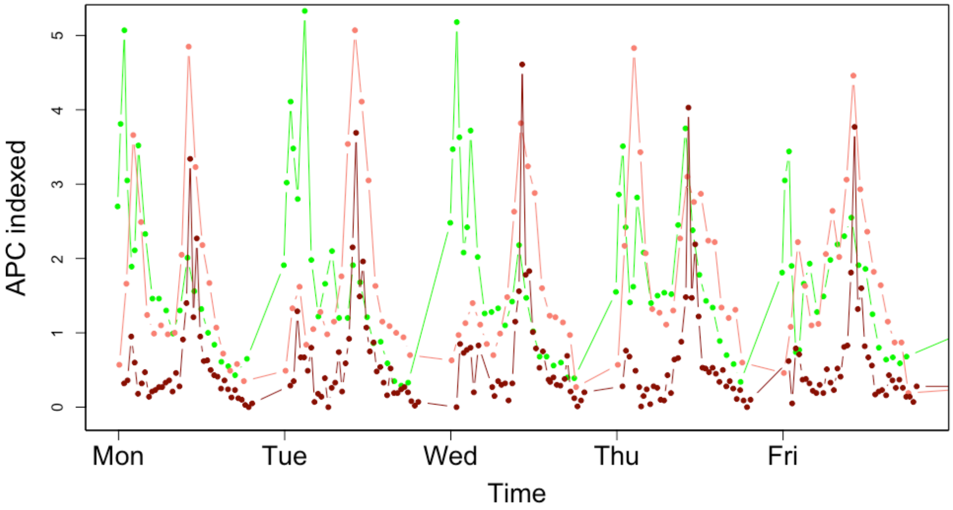

Figure 13. Automatic passenger count, represented as indexed values on the y-axis, for two different train lines as the trains pass base station B5 located before station Y, where green line is Line 1 direction towards the city, light red line is Line 1 away from the city and dark red is Line 2 away from the city.

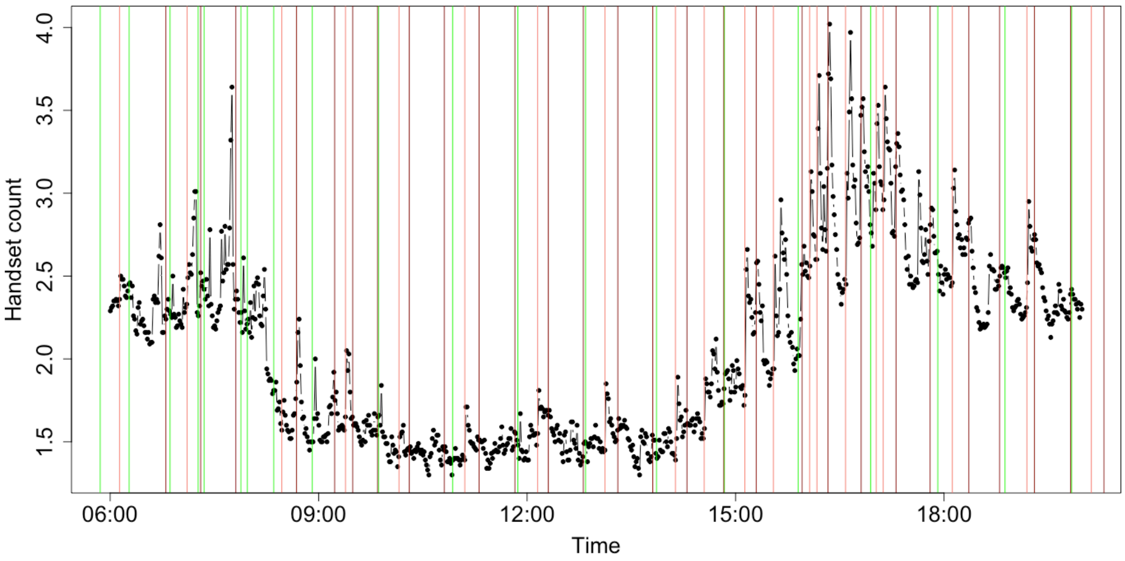

Figure 14. Handset counts from the third data set with one-minute collection time intervals collected on a Thursday at base station B5 located before station Y. The vertical lines represent the same train lines as in Figure 13, where green lines are Line 1 direction towards the city, light red lines are Line 1 direction away from the city, and dark red lines are Line 2 direction away from the city.

The ratios of the number of handsets to the number of travellers (handset count/APC count) are given in Figure 15 for base station B5 from the third data set with one-minute collection time intervals. Figure 15 gives the calculated ratio for trains travelling towards the city and away from the city for five days. A smooth curve was fitted by loess regression as shown by the black line, with confidence bands. For the trains travelling towards the city (see Figure 15 (b)) the ratio centres around one from 6 a.m. in the morning to 3 p.m. in the afternoon. From 3 p.m. when the evening rush hours commence the ratio increases towards around three and four. Ratio above 1 means that the handset counts are higher than the automatic passenger counts. For the trains travelling away from the city (see Figure 15 (a)) the ratio has a close to inverted shape, with a decrease from around four in the morning from 6 a.m. to 10 a.m. during the morning rush hours. From 10 a.m. the ratio has a small curve centred around one, and with a slight increase towards the evening. The ratio for trains travelling towards the city on Thursday and Friday in Figure 15 (b) has three higher ratio-values on, respectively, one and two of the morning trains. For trains travelling away from the city in Figure 15 (a), four of the morning trains on Thursday has lower ratio-values than the other days.

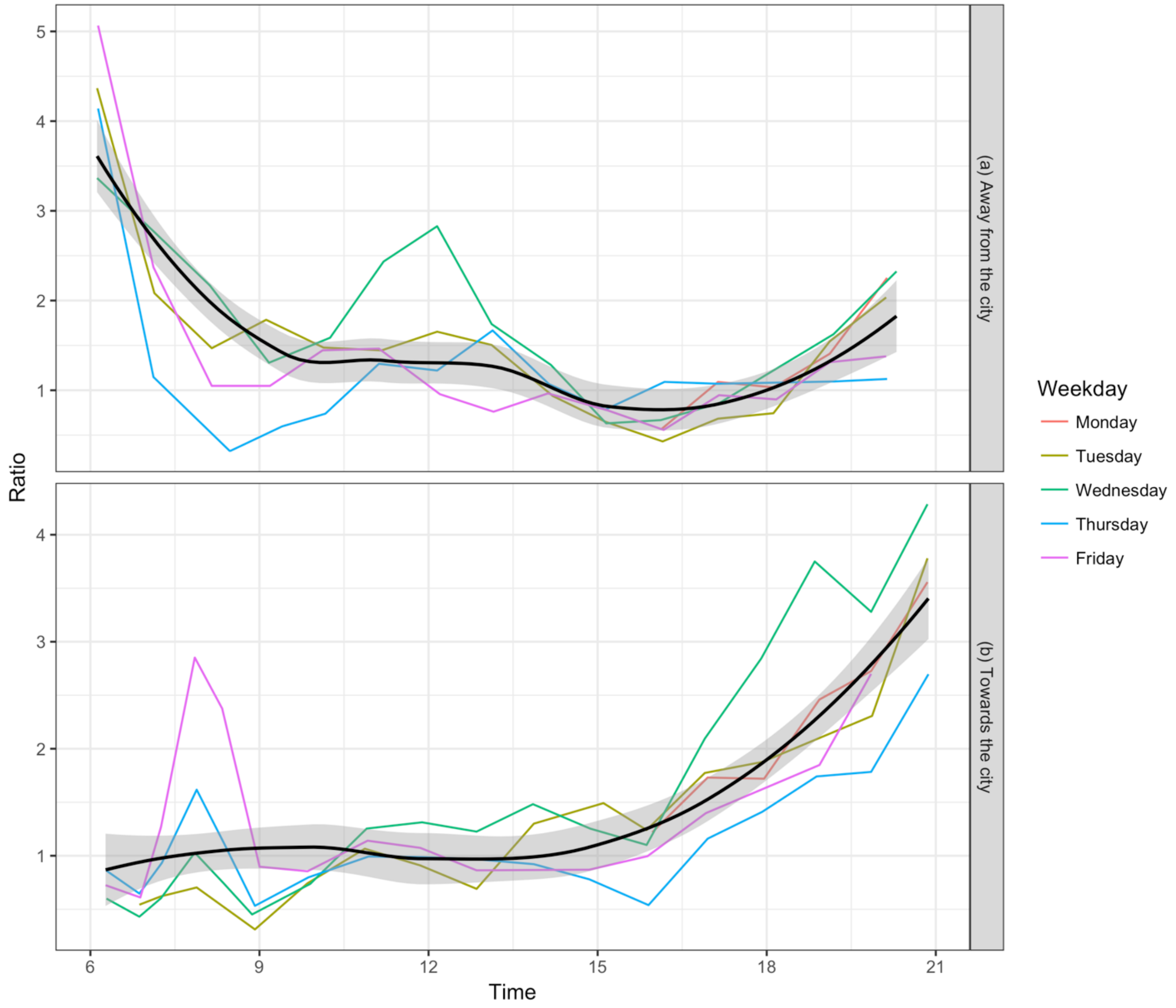

Figure 15. Ratio of handset counts to the number of passengers (APC) for trains going towards the city (b) and away from the city (a) on Line 1 for five days at base station B5 located before station Y. The black lines represent a smooth curve fitted by loess regression with confidence bands.

5 DISCUSSION

This section discusses some of the results presented in Section 4.

5.1 Graphic inspection and collection time interval

The expected relationship between the rush hours and the direction (see assumption 3 and 4 in Section 3.4.4) was not confirmed in Figure 6. However, in Figure 7 (a), we noticed that for base station B5 with one-minute collection time intervals, a large share of the peaks is in connection with trains travelling away from the city in the afternoon rush hours, which is what we were expecting. One-minute collection time intervals appear as more favourable than the intervals of five minutes both in capturing the frequency in peaks and peaks that are connectable to trains passing the base station, as seen in Figures 5, 6 and 7.

Station Y has higher train frequency than the previously studied stations, i.e., station U and station V, as seen in Figure 7. Several of the peaks in the handset counts corresponds nicely with trains that pass the base station. Some of the peaks have collection times between two different train passing times. In these cases, the train passing related to the peak is most likely the time prior to the peak, because the mobiles are registered on the base station until they are connected to a new base station, so the mobiles are most likely registered on the selected base station also for a small amount of time after the train has passed. This is the logic behind the suggested algorithm. However, the question is if this also applies when the collection time is closer to the time of the train that passes right after the collection time. When the trains pass the base station with a high frequency and the time interval between trains are small some of the peaks seem to be placed in such a way that we cannot be certain which peak corresponds with which train.

5.2 Statistical analysis

The violin plots were used in Section 4.3 to analyse the output of the algorithm. The median value was slightly higher for the 'Yes' categories (see Figure 8). For the morning rush hours, the median value was highest for trains travelling towards the city, supporting the assumption made in Section 3.4.4 (assumption 3). For the evening rush hours, the median value was highest for trains travelling away from the city, in agreement with the assumption (assumption 4). Figure 8 showed bimodal distributions for both 'Yes' and 'No' categories with two local maxima in each of the density functions (i.e., for the categories morning rush hours and rest of the day). The low count values in the 'Yes' category was expected because the graphic inspection has shown that there are not peaks every time a train passes (see for instance Figures 7 and 14). What is more concerning is that the 'No' category has so many high values. Figure 9 showed an average daily variation with distinct variation between the morning, the middle of the day and the evening. Thus, this may be a reason why the 'No' category in Figure 8 has a high density of handset counts in both high and low count values. Another explanation could be that some peaks are caused by trains that are about to pass, but are a few seconds before the collection time. Hence, the high value in handset count would be categorised as a 'No' category by the algorithm.

5.3 Extracting the peaks in handset count

The violin plots of the percentage increase (Figure 12) show that the 'No' category has values with positive increases, and also increases of more than 10%. These peaks are probably deviations that are not caught by the algorithm, as discussed above.

5.4 Comparison to actual ridership

The shapes of the ratio curves in Figure 15 was expected because of the daily variation of the handset counts and the APC data. This is illustrated in Figure 16, showing the assumed daily variations based on the previously presented reslutls of handset counts, APC counts and ratio. The daily variation in the handset count (see Figures 9 and 14) show low values in the middle of the day and higher values before and during the morning rush hours and during and after the evening rush hours. First inspecting the trains travelling towards the city, we see that the ratio increases in the evening (see Figure 15 (b)) because the handset counts increases towards the evening (Figure 14) while the APC data remain low (Figure 13). Likewise, for the trains travelling away from the city we see that the ratio starts at a high value and decreases because the handset counts have high values in the morning and then decrease (Figure 14) while the APC data is low from the morning until the evening rush hours commence (Figure 13).

For trains travelling towards the city on Thursday and Friday, the higher ratio-value on three of the morning trains (see Figure 15) are possibly because the APC data show fewer passengers on these days in the morning compared to the other days (see Figure 13). In the same way, for trains travelling away from the city on Thursday, the lower ratio-values on four of the morning trains (see Figure 15) are probably because the APC data show more passengers on this day in the morning compared to the other days (Figure 13).

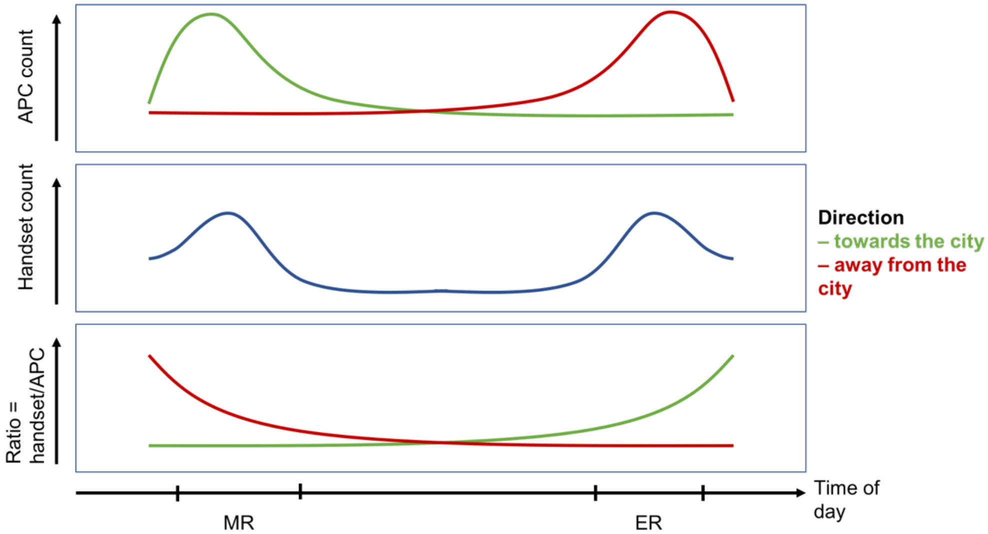

Figure 16. Illustration of the logical connection between the handset counts, the APC data and the ratio. The drawn curves are the supposed daily variation in APC count, handset count and ratio, respectively. The assumed daily variations are based on the results presented in Figures 9 and 14 (handset count), Figure 13 (APC), and Figure 15 (ratio handset count/APC).

6 CONCLUDING DISCUSSION

This study has investigated the potential for using mobile phone data to describe travel patterns that include train travel. We have tested the use of mobile phone data to measure train ridership. We find that there is a connection between the train passing and changes in the handset counts. Although the variation is different for different base stations, there is a significant positive increase when trains pass.

We have shown that it is possible to combine mobile data with railway infrastructure and train traffic data. These preliminary results show that there is a connection between the train passing and changes in handset counts. However, it is also evident that a lot of the trains passing the base station are not detectible in the handset counts. Furthermore, some peaks seem to occur even though there were no trains passing the base station. We also showed that one-minute collection time intervals are needed, especially on the railway sections that have high frequency of trains passing.

The main implication of the findings is that mobile phone data can potentially be used for ridership analyses.

6.1 Combining mobile phone data with railway infrastructure and train traffic data (RQ1)

This study show that it is possible to combine mobile phone data with railway infrastructure and train traffic data, thus answering the first research question. We successfully combined data on rail traffic, timetables and infrastructure with handset counts from mobile phone base stations. Provided that the mobile phone base stations are chosen carefully and the collection time intervals for handset counts are suitable, the handset counts correspond well with the passing of trains. We tested collection time intervals of five minutes and one minute. Especially for high frequency parts of the railway network, a one-minute collection time is needed. Even for lines with less traffic, a one-minute data set appears to have advantages over a five-minute data set.

However, the findings also show that there is potential to improve the method of connecting the collection times with the times of trains passing the base station. As Figure 8 showed, the distribution of handset counts that are not connected to a train passing shows a larger portion of high values than we would prefer. This could indicate that the algorithm is not able to connect all the peaks in handset counts to trains passing. For instance, the base station has a certain range, so considering the time frames in which the trains are passing within the range of the base station and how early or late in that time frame, it is likely that the mobile phones are connected to the base station in question. And we could consider the extension of the trains from the first to the last carriage, which can vary.

To consider the range of the base station, a time margin can be included in the method to connect collection times with when trains pass the base station, which can be done in two ways. Either time can be added on the *train_time*, some before and some after the calculated point in time when the train is passing the base station, to include when one thinks the train is within range of the base station. Depending on how large this time interval is, a result can be that the train will be connected to more than one collection time. This does not necessarily have to be a disadvantage. Alternatively, one can add a time interval in the collection time, for instance, by saying that if the train's time falls within the interval created as a half minute before and a half minute after the collection time, then the train is connected to the collection time. A third way could be to both add a time interval around the collection time and a time interval around the train time. In this approach, the result would most likely be that the trains will be connected to more than one collection time.

6.2 Suitable formats for presentation (RQ2)

One research question was what suitable formats are for presenting and analysing train ridership based on mobile data. We have tested and described different formats for analysing and presenting train ridership based on handset counts. In particular, we have used both absolute numbers of handsets and changes in the handset counts. We have also tried different approaches for estimating the probability that a count peak is related to the passing of a train. Although it is possible to draw conclusions and validate assumptions from graphic analysis, there is a need to develop a uniform mechanism for categorizing the available data and assigning as well as calculating variables for validating the assumptions. Hence, an algorithm was developed yielding the required output variables so that tangible results in relationship to the hypothesis can be extracted for the categories both in preliminary and time and direction analysis.

This section discusses the strengths and weaknesses of each of the formats used in this paper for presenting and analysing the results.

6.2.1 *Graphic inspection and collection time interval*

Graphic inspection was used in Section 4.1 to visualise the handset count and compare the collection times to the calculated approximate times of when trains are passing the base station. The strength of graphic inspection is that it is good to visualise how the collected handsets and trains are connected. The weakness of graphic inspection is that it is less suitable to comparing several days or base stations.

6.2.2 *Violin plot*

The strengths of the violin plot are that it allows us to compare and visualize the relationship between numerical and categorical variables, as well as showing the distributional characteristics of the data. The violin plots would be suitable for comparing distributions of handset counts (and percentage increases) in a study to compare several methods, for instance, to test different methods of calculating the times the trains are passing the base station or for testing other methods of connecting train times with collection times, as discussed above. What is desired is that the category that contains collection times when no trains passed has no high peaks in handset counts. As Figure 9 showed, it is an advantage to separate the handset counts in rush hours and the rest of the day with no rush hour. As the APC data show, more people are travelling towards the city in the morning and away from the city in the evening than during the rest of the day. Thus, even though the violin plots do not show a clear difference between these categories, it will be favourable to make this separation between directions in further testing.

6.2.3 *Extracting the peaks*

As learned from the results on collection time intervals, one-minute collection intervals give better precision in detecting peaks that occur in close time proximity to trains passing the base station. The five-minute collection time interval was less accurate. Extracting peaks on a data set collected at one-minute collection time intervals could be more useful than with five-minute intervals as we saw from the results in Section 4.4.

The violin plots of the percentage increase on the data set with one-minute collection time interval showed that both high and low values were connected through the algorithm to trains passing the base station. And plots of the handset count like Figure 14 show that handset count seems to increase in stages, or at least that as a train passes the handset count increase gradually in more than one collection time. The difference in handset count may therefore be questioned as a good approach to extract the peaks. A method similar to SMA could perform better when the collection time intervals are less than five minutes.

6.2.4 *Comparison to APC ratio*

A useful way to utilise the handset count data would be to find an average ratio between the handset count and the automatic passenger count. The comparison of the actual passenger counts with the number of handsets needs to take specific direction and timeframe into consideration to gauge the exact number of travellers on the passing train. Thus, to calculate the actual number of passengers on the train, the ratio of numbers of local people connected to the base station is one of the factors that needs to be taken into account. The ratio is almost constant during specific times. Further testing of how to connect the train times to the collection times will also affect the ratio between the APC data and handset count. A good connection between the train times and collection times will improve the results of the ratio and will facilitate the ratio serving as a practical validation indicator to observe the effect of passengers on the train with the number of counts.

6.3 Measuring number of passengers using mobile phone data (RQ3)

The last research question was to what extent the format of available mobile phone data is suitable for measuring the number of mobile units passing close to the railway line. The mobile phone data available for this study was handset counts. The handset counts are compared to the number of travellers as measured by on-board APC equipment. The results indicate that handset counts and changes in handset counts can give an indication of the number of travellers on a passing train. The results are, however, not as conclusive as for train detection. The ratio of handsets over passengers varies. It is likely that a large-scale calibration is needed, using more data than we had available, to increase the accuracy of handset counts as indicators of the number of travellers.

To find a good approximation of the number of passengers on the train, we should consider factors that can affect the precision of the measurement. There are some issues that may introduce bias into the handset counts. For instance, there are most likely some people without phones on the trains, or possibly business people with more than one phone. Another issue is that these data are only from one telecom operator, and the number of travelling passengers can be unpredictably divided among different mobile operators. Other factors that can affect the precision of the measurement include,

- the daily variation (morning-afternoon-evening);
- the need to gather data for longer time periods. For evaluation purposes, there is a need for data covering a relatively long period, typically a few years. We recommend that data are stored with the highest possible resolution and that how the data are collected and processed is clearly described. In addition, it would help to find better average values, maybe over years, for monthly variations. Mobile data for longer time durations will account for seasonal effects and smooth the effect of irregular fluctuations in the data and can provide a better bias for developing more precise mathematical model;
- if available, base stations located in railway tunnels would be preferred; and
- taking into account the normal variations of mobile data and specific factors and conditions associated with the specific cell sites can provide a more linear relationship with the ridership and closeness of cell sites to the train station (indoor cell sites for the stations or located exactly at the train station). Variations between the base stations include different areas, different prerequisites for the base stations to pick up phone signals, or numbers of people passing or being in the area.

Other uncertainties with this approach of measuring train ridership are that we cannot say for sure that the time stamps for the collection times and the train times are taken by synchronized watches. There may be a small time-lag that we do not know about.

When discussing the suitability of the available mobile phone data it is also reasonable to disucss whether it is worthwhile to invest and to apply mobile phone data for ridership analysis. If universial and scalable methods are developed, then the costs should be significantly less than with methods used now.The data is already there, so a lot of the expences lies in developing systems and testing good algorithms.

Table 6 summarises some pros and cons of handset counts and mobile phone data.

Table 6. Pros and cons of handset counts and mobile phone data.

Pros	**Cons**
Need no direct access to train or railway network – the source of data and data owner is the mobile network operators	Little experience
Preserves personal privacy (for handset counts)	Personal privacy (not for handset counts, but an issue for other mobile phone data)
Data is available	Separate logs for each mobile network (this study is with mobile phone data from one network)
The data gives units connected to base station, (and potentially movement in network)	The ultimate destination is not provided – at this point (which is true for the data source utilised in this paper, however other mobile phone data may provide destination information)
The user role in data collection previously achieved by the survey process is minimized	The data cannot provide information on trip purpose or on user assessment of service
Improved data quality and increased amount of statistics available	Costs regarding developing systems, testing good algorithms and safeguarding personal privacy.

6.4 Further research

The statistical and graphic analyses and the evaluation of hypothesis developed provide a clear picture of different aspects that can be analysed and looked into combining mobile data and train data. The important fact is that this is only one format of data that has been looked into in detail. The format provides a good starting point for looking into the concept of utilizing mobile data in ridership evaluation and that longer data duration and higher resolution can point out the exact quantifiable figures for the number of travellers. Other formats of anonymous data formats also need to be explored separately to determine if they can be combined with the format used in this report. The availability of longer time interval data and other formats can provide the opportunity of developing systemic models and frameworks for analysing the ridership from different angles. Hence, mobile data can be declared a viable source for calculating the number of travellers on trains.

Further steps needed to make mobile phone data usable tools for measuring ridership on trains:

- Calibrate the ratio of handset counts to the number of passengers using data sets for longer time periods
- Test other methods to determine if a train is connected to the base station at the moment of a collection time.

The easiest approach to measure the number of passengers would be to get direct access to APC and other ridership data. However, mobile phone data opens up a wider set of options for analysis, provided that personal privacy issues are managed and the telecom operators can develop business models for supplying such data.

7 ACKNOWLEDGEMENTS

We thank the Concept program for support and constructive feedback on previous versions of parts of this text.

8 REFERENCES

Alexander, L., Jiang, S., Murga, M. & González, M. C. (2015). Origin-destination trips by purpose and time of day inferred from mobile phone data. *Transportation Research Part C: Emerging Technologies*, 58, Part B, 240-250.

Assemi, B. Safi, H. Mesbah, M. & Ferreira, L. (2016) Developing and Validating a Statistical Model for Travel Mode Identification on Smartphones. IEEE Transactions On Intelligent Transportation Systems, 17(7).

BaneNor (2017). Network Statement. BaneNor. http://www.banenor.no/en/startpage1/Market1/Network-Statement/. Accessed December 6, 2017.

Becker, R., Cáceres, R., Hanson, K., Isaacman, S., Loh, J. M., Martonosi, M., Rowland, J., Urbanek, S., Varshavsky, A. & Volinsky, C. (2013). Human mobility characterization from cellular network data. *Communications of the ACM,* 56**,** 74-82.

Bianchi, F. M., Rizzi, A., Sadeghian, A. & Moiso, C. (2016). Identifying user habits through data mining on call data records. *Engineering Applications of Artificial Intelligence,* 54**,** 49-61.

Boyle, D. B. (1998). Passenger counting technologies and procedures. TCRP Synthesis of Transit Practice 29. Washington, DC: Transportation Research Board.

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

Calabrese, F., Di Lorenzo, G., Liu, L. & Ratti, C. (2011). Estimating Origin-Destination Flows Using Mobile Phone Location Data. IEEE Pervasive Computing, 10(4), 36-44. http://dx.doi.org/10.1109/mprv.2011.41

Chaudhary, M., Bansal, A., Bansal, D., Raman, B., Ramakrishnan, K. K. & Aggarwal, N. (2016). Finding occupancy in buses using crowdsourced data from smartphones. In *Proceedings of the 17th International Conference on Distributed Computing and Networking*. 35. ACM. doi: 10.1145/2833312.2833460

Chu, K. A. & Chapleau, R. (2008). Enriching Archived Smart Card Transaction Data for Transit Demand Modeling. *Transportation Research Record: Journal of the Transportation Research Board*. 2063, 63–72. doi: 10.3141/2063-08

Doi, M. & Allen, W. B. (1986). A time series analysis of monthly ridership for an urban rail rapid transit line. *Transportation*, 13(3), 257–269. https://doi.org/10.1007/BF00148619

Drageide, V. (2009). Towards Privacy Management of Information Systems. MS thesis. The University of Bergen.

Fowkes, A.S., Nash, C.A. & Whitening, A. E. (1985) Understanding trends in inter-city rail traffic in Great Britain. *Transport Planning and Technology*, 10, 65-80

Frias-Martinez, V., Soguero, C. & Frias-Martinez, E. (2012). Estimation of urban commuting patterns using cellphone network data. In *Proceedings of the ACM SIGKDD International Workshop on Urban Computing*, 9-16.

Gordillo, F. (2006). *The value of automated fare collection data for transit planning: an example of rail transit od matrix estimation* (Thesis, S. M., Massachusetts Institute of Technology). http://hdl.handle.net/1721.1/38570

Gundlegård, D., Rydergren, C., Breyer, N. & Rajna, B. (2016). Travel demand estimation and network assignment based on cellular network data. *Computer Communications*, 95, 29-42.

Higuchi, T., Yamaguchi, H., Higashino, T. (2015). Mobile devices as an infrastructure: A survey of opportunistic sensing technology. *Journal of information processing*, 23(2), 94-104.

Hintze, J. L. & Nelson, R. D. (1998). Violin plots: a box plot-density trace synergism. The American Statistician, 52, 181-184.

Holleczek, T., Yu, L., Lee, J. K., Senn, O., Ratti, C. & Jaillet, P. (2014). Detecting weak public transport connections from cellphone and public transport data. Proceedings of the 2014 International Conference on Big Data Science and Computing, 2014. ACM, 9.

Järv, O., Ahas, R., Saluveer, E., Derudder, B. & Witlox, F. (2012). Mobile phones in a traffic flow: a geographical perspective to evening rush hour traffic analysis using call detail records. PLoS One, 7(11), e49171. doi:10.1371/journal.pone.0049171

Jiang, S., Yang, Y., Fiore, G., Jr., J. F., Frazzoli, E. & González, M., (2013). A review of urban computing for mobile phone traces: current methods, challenges and opportunities. In *Proceedings of the ACM SIGKDD International Workshop on Urban Computing*.

Kujala, R., Aledavood, T. & Saramäki, J. (2016). Estimation and monitoring of city-to-city travel times using call detail records. *EPJ Data Science*, 5(1). doi:10.1140/epjds/s13688-016-0067-3.

Larijani, A. N., Olteanu-Raimond, A.-M., Perret, J., Brédif, M. & Ziemlicki, C. (2015). Investigating the mobile phone data to estimate the origin destination flow and analysis; case study: Paris region. *Transportation Research Procedia*, 6, 64-78.

Leo, Y., Busson, A., Sarraute, C. & Fleury, E. (2016). Call detail records to characterize usages and mobility events of phone users. *Computer Communications*, 95, 43-53. doi:10.1016/j.comcom.2016.05.003

Li, J. P. (2000). Train station passenger flow study. In *Proceedings of the 2000 Winter Simulation Conference.*

Mac Donald, V. H. (1979). Advanced Mobile Phone Service: The Cellular Concept. *Bell System Technical Journal*. 58(1), 15–41. doi:10.1002/j.1538-7305.1979.tb02209.x

Nielsen, B. F., Frølich, L., Nielsen, O. A. & Filges, D. (2014). Estimating passenger numbers in trains using existing weighing capabilities. *Transportmetrica A: Transport Science*, 10(6), 502-517.

Olsson, N. O., & Bull-Berg, H. (2015). Use of big data in project evaluations. *International Journal of Managing Projects in Business*, 8(3), 491-512.

Pelletier, M.-P., Trépanier, M., Morency, C. (2011). Smart card data use in public transit: A literature review. *Transportation Research Part C: Emerging Technologies*, 19(4), 557-568. doi:10.1016/j.trc.2010.12.003

Schoeni, R. F., Stafford, F., Mcgonagle, K. A. & Andreski, P. (2013). Response Rates in National Panel Surveys. *The Annals of the American Academy of Political and Social Science*. 645(1), 60-87. doi:10.1177/0002716212456363

Sørensen, A. Ø., Olsson, N., Akhtar, M. & Bull-Berg, H. (2017). *Approaches, Technologies and importance of Analysis of Number of Train Travellers*. Manuscript submitted for publication.

Teknisk Ukeblad (2014). *Matematiske verktøy skal gi kortere stop på togstasjonene*, January 29 2014, http://www.tu.no/artikler/matematisk-verktoy-skal-gi-kortere-stopp-pa-togstasjonene/224974

Vigren, A. (2017). Competition in Public Transport. Essays on competitive tendering and open-access competition in Sweden. Doctoral thesis in transport science. KTH Royal Institute of Technology, Stockholm, Sweden.

Vuchic (2005). *Urban transit. Operations, planning and economics.* John Wiley and Sons, Hoboken, NJ.

Wang, W., Attanucci, J. P. & Wilson, N. H. M. (2011). Bus Passenger Origin-Destination Estimation and Related Analyses Using Automated Data Collection Systems. *Journal of Public Transportation*, 14 (4): 131-150.

Xu, Y, Shaw, S-L, Fang, Z & Yin, L (2016). Estimating Potential Demand of Bicycle Trips from Mobile Phone Data – An Anchor-Point Based Approach. *ISPRS International Journal of Geo-Information*, 5(8), 131; doi: 10.3390/ijgi5080131.

Zhao, J., Rahbee, A. & Wilson, N. H. M. (2007). Estimating a Rail Passenger Trip Origin-Destination Matrix Using Automatic Data Collection Systems. *Computer-Aided Civil and Infrastructure Engineering*, 22(5), 376-387.

Zhao, Z., Shaw, S.L., Xu, Y., Lu, F., Chen, J. & Yin, L. (2016). Understanding the bias of call detail records in human mobility research. *International Journal of Geographical Information Science*, *30*(9), 1738-1762.