Transportation
https://doi.org/10.1007/s11116-020-10151-7

Mobile phone data in transportation research: methods for benchmarking against other data sources

Andreas Dypvik Landmark[1] · **Petter Arnesen**[2] · **Carl-Johan Södersten**[2] · **Odd André Hjelkrem**[2]

Abstract
The ubiquity of personal cellular phones in society has led to a surging interest in using Big Data generated by mobile phones in transport research. Studies have suggested that the vast amount of data could be used to estimate origin–destination (OD) matrices, thereby potentially replacing traditional data sources such as travel surveys. However, constructing OD matrices from mobile phone data (MPD) entails multiple challenges, and the lack of ground truth hampers the evaluation and validation of the estimated matrices. Furthermore, national laws may prohibit the distribution of MPD for research purposes, compelling researchers to work with pre-compiled OD matrices with no insight into the methods used. In this paper, we analyse a set of such pre-compiled OD matrices from the greater Oslo area and perform validation procedures against several sources to assess the quality and robustness of the OD matrices as well as their usefulness in transportation planning applications. We find that while the OD matrices correlate well with other sources at a low resolution, the reliability decreases when a finer level of detail is chosen, particularly when comparing shorter trips between neighbouring areas. Our results suggest that coarseness of data and privacy concerns restrict the usefulness of MPD in transport research in the case where OD matrices are pre-compiled by the operator.

Keywords Mobile phone data · Origin–destination (OD) estimation · Travel surveys

Introduction

Origin–destination (OD) matrices are a crucial component in the design of transportation networks and the development and planning of public transportation. Traditionally, OD matrices have been constructed with the help of household questionnaires and road surveys. Such sources constitute an invaluable source of information as they provide trip details at a fine detail (origin and destination, recurrence of trips, transport mode, trip purpose, etc.) and can be designed to answer precise research questions. In Norway, the public

✉ Carl-Johan Södersten
carl.sodersten@sintef.no

[1] Department of Technology Management, SINTEF, Trondheim, Norway

[2] Department of Mobility and Economics, SINTEF, Trondheim, Norway

Published online: 02 January 2021

transport authority for the capital metropolitan area performs around 9000 interviews each year in order to map and understand travel habits of inhabitants in Oslo and Akershus (joint population of 1.3 million). The outcome of the interviews can then be used to estimate the number of trips undertaken between predefined areas of the cities so that OD matrices can be constructed.

There is, however, an inherent drawback associated with individual surveys. The preparation, collection and processing are costly and time-consuming. In a 2006 study of household travel surveys, Stopher and Greaves (2007) estimated that the costs of conducting a face-to-face travel survey amounted to $350 per household. Furthermore, sample sizes are small and keep decreasing; surveys often cover much less than 1% of the population (Stopher and Greaves 2007). Hence, the collected data needs to be considerably scaled up to generalise for the entire population. Increasing costs and safety concerns have led to a shift towards telephone-based interviews. These have been shown to be less accurate as respondents tend to omit a significant number of trips (between 20 and 30% for most surveys) (Forrest and Pearson 2005; Wolf et al. 2003) and the self-reported answers are not always reliable because respondents of such surveys often tend to over- or underestimate their own consumption (in this case of transportation). While face-to-face interviews are reportedly more accurate and complete, Stopher et al. (2007) still find a shortfall of trip recording between 7 and 12%. Household travel surveys also suffer from a low response rate, which tends to decrease as the size of the household and travel frequency increase, since these factors typically make the completion of a travel survey more burdensome.

The demand for more extensive, accurate and complete travel data has led to the emergence of alternatives to these traditional data collection methods (Stopher and Greaves 2007). The use and ubiquity of mobile phones have exploded in the last 2 decades, making the prospect of using data signals from mobile phones in traffic applications interesting, particularly as a data source for constructing OD matrices. Dozens of studies applying mobile phone data (MPD) to estimate vehicle and passenger flows have been published, and a lot of the research has focussed on the development of algorithms that filter through the vast amount of data generated in an attempt to identify travel patterns, in many cases in large cities such as Singapore (Aksehirli and Li 2018; Holleczek et al. 2015; Poonawala et al. 2016), Boston (Alexander et al. 2015; Qu et al. 2015; Wang et al. 2010; Toole et al. 2015; Vazifeh et al. 2019; Calabrese et al. 2011a), Paris (Bachir et al. 2019; Aguiléra et al. 2014; Smoreda et al. 2013; Larijani et al. 2015) and Barcelona (Bassolas et al. 2019; Montero et al. 2019). In recent years, access to mobile phone data has become a commodity in many markets and is often available to public authorities (e.g. transport planners) for use in non-research/commercial purposes. The data in these studies range from simple counts (with temporospatial filtering) and various forms of aggregated data to OD matrices. However, national laws may restrict the resolution and detail of the distributed data. In Norway, privacy laws prohibit the distribution of raw MPD to transportation researchers. As such, it is currently only possible to purchase OD matrices that have been pre-compiled by the mobile phone operators. The pre-compilation procedure includes anonymisation of the data as well as additional censoring measures that prevent potential re-identification of users.

In this paper, we aim to assess whether a set of OD matrices that have been compiled by a mobile phone operator, with little to no insight into the compilation procedure given to the end user, could be used as a potential replacement for travel surveys in transport planning applications, for instance by a public transport management company. We begin with an analysis of the OD matrices purchased from a mobile phone operator, describing the main challenges involved in their construction and discussing an optimal spatial and

temporal resolution. Since the OD matrices were computed, processed and anonymised by the mobile phone operator prior to being distributed, we begin with several preliminary assessments of their reliability. We then proceed with comparisons of the matrices with OD matrices constructed from other sources such as population statistics, ridership counts and traditional travel surveys, and then discuss on the usefulness of these pre-compiled OD matrices.

Background

Given the described drawbacks of traditional data collection methods, MPD has emerged as an appealing alternative for constructing OD matrices. The pervasiveness of mobile phones in today's society makes it an optimal source for obtaining data from a large share of the population [worldwide mobile subscription rate now exceeds one device per person (World Bank 2019)] and covering a vast spatial scale in almost all regions of the world. As computational power is constantly increasing, the magnitude of the data becomes less of the impediment it once was, allowing the big data generated by mobile phones to be processed relatively quickly, which makes MPD a tempting replacement for traditional data sources such as surveys or data logs from electronic travel cards. This has led to a wave of studies assessing the usefulness of MPD in transport research, many of which concluding on a positive note. Nevertheless, a range of difficulties has been highlighted.

Multiple papers mention the spatial accuracy and coarseness of MPD as a limiting factor (Ahas et al. 2010; Calabrese et al. 2011a, 2013; Gundlegård et al. 2016; Huang et al. 2019; Montero et al. 2019; Di Lorenzo et al. 2015; Becker et al. 2013; Bachir et al. 2019; Phithakkitnukoon et al. 2017; Poonawala et al. 2016; Aksehirli and Li 2018; Wang et al. 2018). Some studies suggest that the highest resolution of MPD ranges from 200 to 400 m in urban areas (Kalatian and Shafahi 2016; Huang et al. 2019), while others find localisation errors of several kilometres (Di Lorenzo et al. 2015). This entails several problems. Firstly, it makes the differentiation between transport modes difficult, e.g. between car drivers and public transport users, for instance when roads and train tracks run parallel (Aguiléra et al. 2014; Bachir et al. 2019; Phithakkitnukoon et al. 2017). Secondly, since antenna density is strongly correlated with population density, the spatial resolution rapidly decreases with the remoteness of the analysed areas (Gundlegård et al. 2016), while overlapping antenna coverage may blur shorter trips in populated areas (Wu et al. 2013). Thirdly, the coarse spatial resolution also limits the minimum size of the regions that can be considered (Calabrese et al. 2011a), and makes it hard to determine what constitutes a trip (Gundlegård et al. 2016). Fourthly, it makes it difficult to detect changes in mode within a trip as well as tracking shorter trips [e.g. biking and walking trips (Bachir et al. 2019)]. Huang et al. (2019) perform a systematic literature review of studies using MPD to detect transport modes and find that only one of the analysed papers differentiates between all available transport modes [which in that specific study include car, bus, tram, train, cycling and walking (Danafar et al. 2017)], but note that the study does not provide any measure of accuracy of the proposed detection algorithm.

Other concerns include the coverage of the data. Most studies are conducted using data obtained from one mobile phone operator, which limits the sample size to the market share of the operator (Calabrese et al. 2011a; Di Lorenzo et al. 2015; Gundlegård et al. 2016; Ni et al. 2018; Sørensen et al. 2018; Doyle et al. 2011; Aguiléra et al. 2014; Schlaich et al. 2010; Chen et al. 2016). As such, sample data is often scaled up using e.g. population

census data (Chen et al. 2016). Furthermore, most studies are based on call detail records (CDR) and/or internet protocol detail records (IPDR), which implies that the data sampling is event-driven (that is, MPD is only recorded when the phone user is actively using the phone for calling, texting or accessing the Internet). Hence, connection patterns of users affect data capture (Calabrese et al. 2011a, 2013; Gundlegård et al. 2016; Huang et al. 2019; García-Albertos et al. 2019; Phithakkitnukoon et al. 2017; Wu et al. 2013; Yamada et al. 2016). Gundlegård et al. (2016) found that users tend to make fewer calls in the mornings, thereby making CDR less reliable for analysing e.g. morning commute patterns. CDR data is therefore also dependent on users' calling plans (Calabrese et al. 2011a, 2013; Di Lorenzo et al. 2015), particularly in countries where unlimited calling/texting plans are less common.

Perhaps the most crucial limitation concerns the validation of MPD since no ground truth data exists (Huang et al. 2018, 2019; Phithakkitnukoon et al. 2017; Chen et al. 2016). In their literature review, Huang et al. (2019) find that few studies evaluate let alone validate their methods, and conclude that the lack of standardised evaluation procedure and the scarcity of results validations make it difficult to assess which mode detection methods work best. Amongst the studies that provide some degree of data validation, several avenues are taken. Methods include cross-checking against log data from travel cards (Bachir et al. 2019; Montero et al. 2019; Aguiléra et al. 2014; Poonawala et al. 2016) and against mode share statistics data from official source (e.g. census statistics) or self-reported data (questionnaires and surveys) (Becker et al. 2013; Calabrese et al. 2011a, b; Phithakkitnukoon et al. 2017; Qu et al. 2015; García et al. 2016). Other studies validate their data through georeferencing (using geographic data of the analysed area and checking if MPD trajectory intersects with pre-defined spatial areas) (Wu et al. 2013; Li et al. 2017; Doyle et al. 2011; García et al. 2016; Holleczek et al. 2015; Horn and Kern 2015), via manual counts along roads (Bassolas et al. 2019; Iqbal et al. 2014) and public transport stations (Holleczek et al. 2015), or by cross-checking models with e.g. observed congestion observations (Huang et al. 2018), traffic sensor data (Wu et al. 2015) or sighting data (cell tower triangulation) (Wang and Chen 2018). The few studies that do validate against actual ground truth data (for instance by having volunteers install tracking software on their devices or share their GPS coordinates (Becker et al. 2013; Isaacman et al. 2011; Zheng et al. 2010; Xu et al. 2010; Asgari 2016)), do so on a sample size usually several orders of magnitude smaller than the studied region.

Using MPD in transport research also entails some inherent problems associated with how it is collected. For instance, some studies mention the issue of users carrying several devices (e.g. personal and work phones) (Calabrese et al. 2011a, 2013; Di Lorenzo et al. 2015; Doyle et al. 2011; Chen et al. 2016), car-sharing (Calabrese et al. 2011a) and non-randomness of users (e.g. larger groups of users travelling together) (Calabrese et al. 2011a, 2013; Di Lorenzo et al. 2015). Other studies mention the sheer size of the data as being a problem in itself (Ahas et al. 2010) and the fact that the data is only stored for a relatively short period of time (Huang et al. 2018). Additionally, collecting MPD involves issues of privacy (Ahas et al. 2010; Sørensen et al. 2018; Smoreda et al. 2013). Data collection and processing may be impeded by mobile phone operators not sharing e.g. location estimation methods and size of network cells (Ahas et al. 2010; Huang et al. 2019). Moreover, because the data needs to be anonymised before it can be used in studies, MPD does not provide semantic information about users (age, income groups, environmental awareness, etc.) or purpose of trip (work, leisure, travel) (Sørensen et al. 2018; Doyle et al. 2011; Calabrese et al. 2011a; Alexander et al. 2015), which means that it cannot be used to study e.g. underlying behaviour mechanisms of households or individuals and therefore

narrows the scope of analyses that can be done (Phithakkitnukoon et al. 2017; Calabrese et al. 2013). For instance, this entails that MPD is not detailed enough to study how various policy measures (congestion charges, public transport subsidies, etc.) affect different types of trips or income groups—information that could be highly relevant for policymakers. Another consequence of privacy regulation is that data may have to be censored to inhibit the probability of potential re-identification of users from anonymised data.

Data and methods

Data

The MPD used in this study was purchased from one of the two largest mobile phone operators in Norway. Between them, these operators account for an 85% share of the national telecom market. The datasets (summarised in Table 1) are based on CDR, IPDR, as well as cell tower switches (which occur when a device is moving and leads to a shift in the cell tower channelling the activity). These comprise a spatiotemporal snapshot handset counts (dataset A), and two sets of OD matrices constructed by the mobile phone operator; one based on the same spatiotemporal window (dataset B) and one more spatially refined (dataset C). Due to Norwegian privacy regulations, details about how the OD matrices were compiled are scarce, including which algorithms were used to identify trips from cell phone data and how these trips were aggregated into OD matrices. The MPD vendor did state that the provided data had been censored using *k*-anonymity (described in "Data anonymisation" section) and subsequently scaled up to the population of the analysed region.

Further data sources used include population statistics from Statistics Norway (the official Norwegian statistical office), turnpike logs from the road toll operator in the region as well as door counts and travel survey statistics from the public transport management company. To enable comparison of these sources to the MPD data, the additional data sources were also scaled up (linearly) to population.

Defining features of the MPD used in the study

Temporal and spatial resolution

The data analysed stems from the four Wednesdays in February 2019 (one of which occurring during Norwegian school holidays). Wednesdays were chosen as they were deemed representative of "normal" workdays. The OD matrices were computed and processed (i.e. smoothened, anonymised and censored) by the mobile phone operator prior to being delivered to the authors of this study and are provided in 1-h windows.

Table 1 Summary statistics for the datasets

	Number of areas	$\mu_{area} \pm \sigma$ (min–max) km^2	$\mu_{popl} \pm \sigma$ (min–max)
Datasets A and B	40	141 ± 188 (2.64–953.88)	$19{,}170 \pm 22{,}131$ (1077–126,841)
Dataset C	2048	2.35 ± 7.27 (0.009–113,4)	653 ± 592 (0–2963)

Norway is divided into 14,000 demographically homogenous basic statistical units (Statistics Norway 2019) (henceforth referred to as "boroughs"), which are the smallest official statistical unit in Norway. These range from under 1 to 2343 km^2 and in population from 0 to 6304. The region of Oslo consists of 40 "districts" that follow public municipality and neighbourhood delimitations. These districts are aggregates of these boroughs and vary substantially in size, ranging from a few units for the inner-city areas to entire municipalities comprising several hundred boroughs, and correspond to public transport markets.

The main dataset (dataset B) used in this study consists of OD matrices describing trips between the 40 districts along with the number of travellers undertaking the described trips. Figure 1 shows the studied region of the Oslo metropolitan area partitioned into 12 market areas and 40 districts. Two districts are excluded from the study (indicated in red in Fig. 1) as they cover public recreation areas without any major roads, making them less interesting from a transport planning perspective.

An additional dataset (dataset C) of more spatially refined OD matrices was also purchased, covering trips to and from each of the six boroughs that constitute Asker city centre to other boroughs. Asker is a minor city but constitutes an important and significant suburb in the metropolitan area. This latter dataset was used to assess the loss of information that occurs as a result of anonymisation algorithms when borough resolution is chosen. The reason for choosing Asker city centre as a subfield of the study was its strategical importance as a regional highway and railway hub. In datasets A and B, Asker city centre is part of the greater municipality of Asker (pop. 60,000).

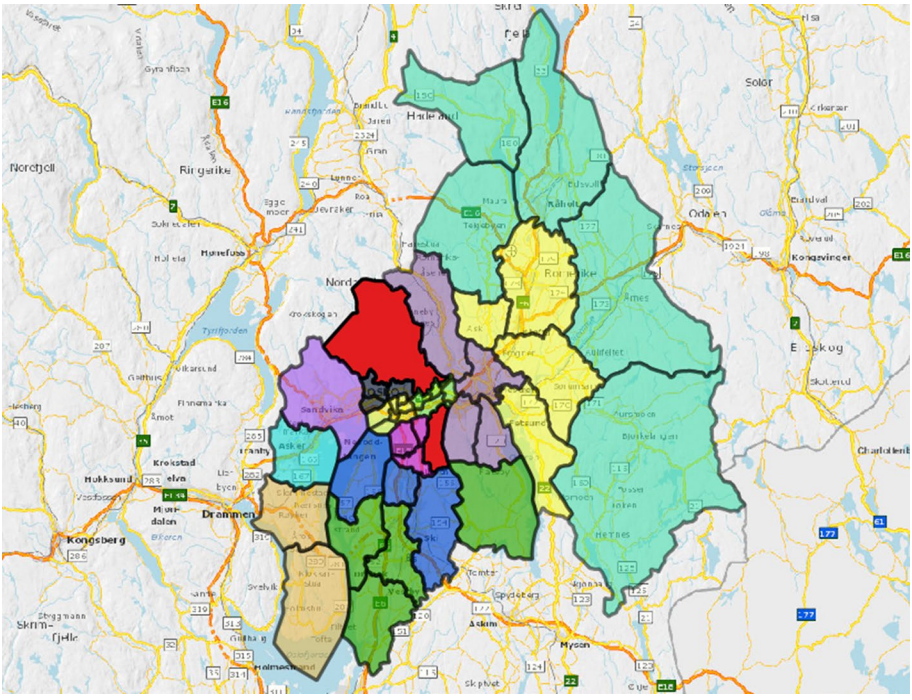


Fig. 1 Division of Oslo metropolitan area into units of study. Transparent colours indicate the 12 market areas defined by the operator, and areas delimited by solid black lines mark the 40 districts. Red areas indicate public recreational/forest areas and are excluded from the study. (Color figure online)

Data anonymisation

Working with data obtained from personal mobile devices entails inherent privacy concerns, and studies using MPD are therefore usually required to adhere to stringent rules to preserve mobile users' integrity and privacy. This includes not only anonymising data but also aggregating data to a level that prevents the reconstruction of individual signals so that the users' (anonymised) travel patterns cannot be re-identified. Further steps may have to be taken depending on national regulations. Norwegian law mandates that CDR should be deleted (or at least anonymised) as soon as they are *no longer needed for billing purposes* [within 3–5 months (Drageide 2009)]. In addition, the law dictates that CDR may only be used for the same specific purpose as they were originally collected for, unless an active consent has been granted by the subscriber (Drageide 2009). As such, using CDR in transport research studies in Norway involves taking additional procedures to ensure that the data cannot be re-traced back to individuals. In practice, this means that the individual device IDs (IMSI) are hashed and assigned a new anonymous identification key each day, thereby making it impossible to track a specific (albeit anonymised) device across multiple days. Secondly, data is only shared if the number of trips between two areas and within a certain timeframe exceeds a certain value k (called k-value threshold or k-anonymity), which the operator has set to 5 (*after* rescaling to population census). This means that if the computed OD matrices result in less than 5 trips being taken between origin O and destination D within the predefined time interval, this value is censored (i.e. set to 0) before the OD matrices are delivered. Such a procedure inevitably leads to a trade-off between spatial and temporal resolution (and ultimately between both variables and data quality/availability); the finer the temporal and spatial resolution, the higher the probability that the number of resulting observations will fall below the required threshold (as illustrated in Fig. 2).

Trip definition

One of the core difficulties of using MPD in the estimation of OD matrices is defining and detecting trips (i.e. trip generation). An advantage with OD matrices based on surveys is that participants can precisely describe their undertaken trips according to the trip purpose, with the change of purpose marking the commencement of another individual trip. However, MPD comprises a set of sampled data points from mobile devices

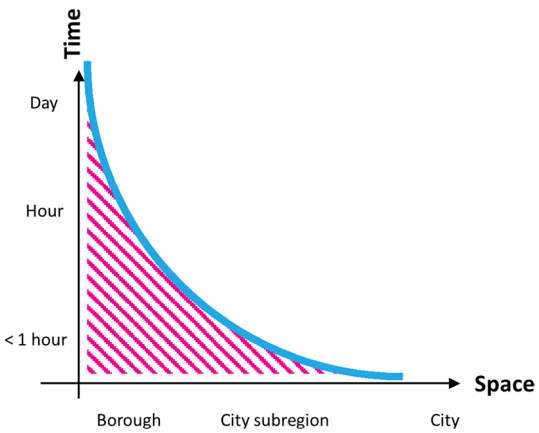


Fig. 2 Conceptual graph illustrating the trade-off between spatial/temporal resolution and data quality that needs to be considered as a result of censoring procedures. The blue line illustrates the k-value threshold, and observations below the line will be censored. (Color figure online)

that need to be processed and analysed in order to distinguish potential trips between different areas, in other words challenging the usual order of trip generation and trip distribution in the traditional transport forecasting method.

As mentioned in the introduction, the spatial resolution of MPD is coarse, and the sampling rate is irregular and depends on user activity, which means that mobile phone signals are unevenly distributed in space and time. The MPD provided by the operator stems from CDR, IPDR and cell tower switches and is therefore event-driven, meaning that unless a user is actively using the device or travelling across different cell areas, data are not sampled. Event-driven data collection has been considered a limitation in previous studies, but since most mobile devices in Norway are almost constantly connected to the internet, this results in a semi-continuous stream of signals being sent to cell towers. According to one operator, a typical mobile phone user in Norway transmits between 400 and 700 signals a day (which corresponds to an average sampling rate of 17–30 signals per hour), though this metric varies substantially depending on user activity, geolocation and time of day. For instance, the sampling rate is lower at night due to less user activity and movement. A lower sampling rate entails that the probability that a short trip goes undetected increases, and it is therefore likely that short trips may be systematically under-represented in the data.

The algorithm used by the operator in this study relies on three main principles to define a trip:

1. A trip is assumed to begin when a previously stationary mobile device switches between network cells
2. A trip is assumed to end when a previously moving device remains stationary (i.e. within one cell) for longer than a certain period (a dynamic threshold)
3. A trip is also assumed to end when a device that has been moving in a certain direction (e.g. undertaking a trip) reverts direction, i.e. when a 180° turn is detected

The threshold described in the second point needs to be exogenously defined. In this study, it is set as a function of trip length, with the rationale that longer trips may include longer stopover times, for instance using public transportation that runs less frequently. The allowed length of a stopover lies between lower and upper bounds of 10 min (for trips under 10 km) respectively 50 min (for trips over 200 km) and increases linearly between these bounds. This threshold function has been set by the mobile phone operator and is embedded in the algorithm used to construct the OD matrices (and could potentially be changed).

The drafting of such principles is necessary to distinguish between trip stopovers and trip destinations, which is, in turn, a prerequisite in the differentiation of the various trips undertaken by the device carrier during the course of a day. Nevertheless, the principles listed above are likely to lead to certain misclassifications. One obvious example would be the case where a particular trip requires a stopover lasting more than 10 min (which is likely to occur during early or late hours of the day when public transport is less frequent). In such a case, the current definitions would imply that a single trip would be reported as two trips (or more). Regarding the third principle, misclassification could occur in the case where a user travels by metro/train until a certain station and then retraces his/her steps above ground to reach the intended destination. This, too, could lead to a single trip being reported as two.

Methods

The principal challenge associated with estimating OD matrices from MPD lies in the processing of the collected data. This entails various pre-processing procedures of filtering out noisy data, data smoothing, sorting, etc., as well as algorithmic segmentation procedures to identify trajectories and/or transport modes, and potentially also post-processing procedures to validate the obtained matrices (Wu et al. 2016). The MPD-based OD matrices used in this study were processed by the mobile phone operator prior to being shared for further analysis due to privacy regulations. As such, we do not have detailed insight into how these matrices were compiled, other than information regarding how trips were defined and how sparse data were censored to prevent potential re-identification of users. We therefore undertook additional validation procedures to assess the reliability of the provided OD matrices, summarised in Table 2. Most of these procedures involve simple mathematical manipulations such as aggregation of data into comparable statistics or statistical calculations such as correlation with other available data sources and are therefore explained in the "Results" section.

We then proceeded with a comparison of the MPD-based OD matrices with OD matrices constructed using public transport data (in "Validation against public transport data" section), turnpike logs (in "Validation against turnpike logs" section) and finally traditional travel surveys (in "Validation against travel surveys" section).

While the comparison to turnpike logs and travel surveys involved simple data aggregation and harmonisation procedures, the comparison to public transport data required converting the available transport logs to OD matrices. The public transportation company keeps count of the number of passengers boarding and alighting from buses and trams running in the greater Oslo area through sensors located at every door. By taking the difference between these two measures, it is possible to estimate the number of passengers travelling along a certain route at a specific time. This, however, does not provide origin and destination points for individual travellers. To derive OD matrices from the passenger counts, a naïve estimation method was applied, where the number of passengers alighting at a certain stop is proportional to the number of passengers boarding. This is illustrated in Fig. 3, where the number of passengers boarding at stop A is assumed to alight at stops B, C and D (final stop) in the same proportion as the *total* number of passengers alighting at these stops. This naïve approach results in estimations for the number of passengers travelling between different stops.

Once all public transport activity had been converted into individual trips, these were reconciled with the 40 studied subregions and temporally aggregated into 1-h periods to obtain comparable harmonised OD matrices.

Table 2 Quality assessment scenarios

Quality assessment scenario	Hypothesis
Comparing different levels of aggregation (in "Effects of data censoring" section)	As a direct result of *k*-anonymity we expect more data to be censored as spatial resolution becomes more refined
Analysing net travels (in "Net travelling balance" section)	Net travels in and out of zones should be close to equal within a 24-h time span
Comparing to population statistics (in "Validation against population statistics" section)	The handset counts in dataset A should correspond to population statistics district-wise

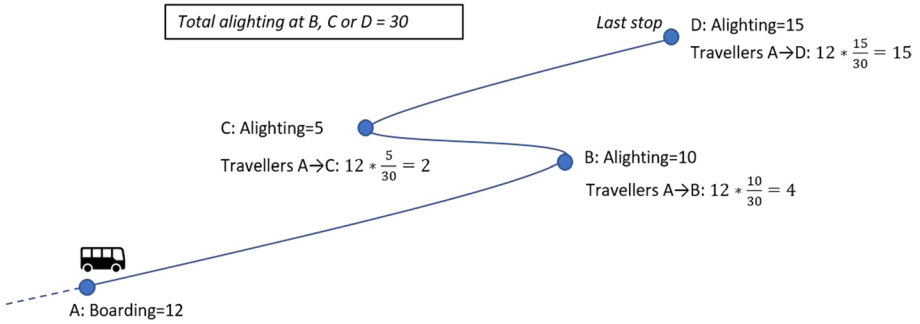


Fig. 3 Illustration of the method used to estimate OD matrices from passenger counts

Results

Effects of data censoring

To assess how the privacy algorithm affects our data, we compared the number of trips recorded in a specific area using district respectively borough resolutions. Figure 4 shows the distribution of recorded OD combinations during a 24-h period over corresponding trip occurrences, for OD matrices based on district (LHS—left-hand side graph) and borough (RHS—right-hand side graph) resolution, respectively. In the LHS graph, we find no occurrences of district-to-district trips below ten. This is because districts are relatively large, and the odds of obtaining few trips between two areas are therefore small. The number of OD combinations then gradually increases to reach a peak at around 20 observations, meaning that 20 trip occurrences are the most common amongst the analysed OD combinations. Consequently, we can safely assume that the privacy algorithm does not

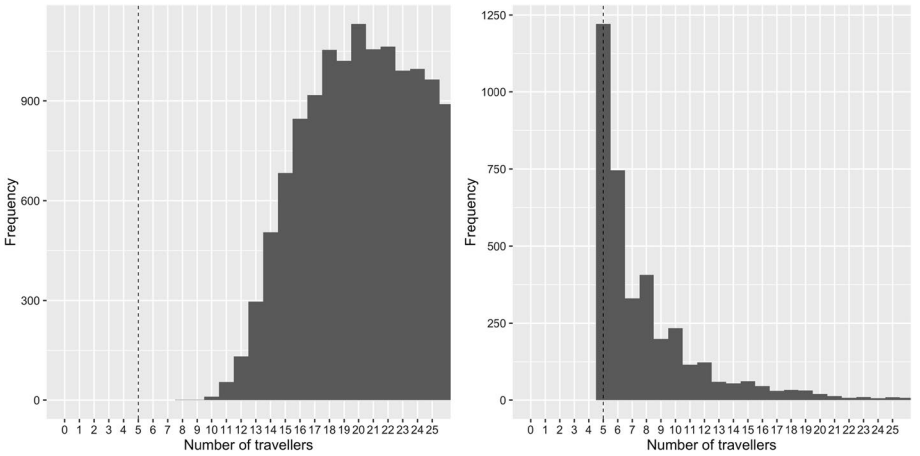


Fig. 4 Histogram of OD matrices at district (left graph) and borough (right graph) resolutions. The vertical axis shows the number of OD combinations that result in the number of trip occurrences on the horizontal axis. Trip occurrences below the k-value threshold of 5 are filtered out (i.e. set to 0)

affect the results when a district resolution is applied, since it is performed independently on each dataset (i.e. at borough respectively district resolution).

The RHS graph, however, peaks at the *k*-value threshold (5 trips), suggesting that more trips would have been observed below the *k*-value had the algorithm not filtered out (censored) those observations (trips). This is an indication that analysing the OD matrices at the borough resolution is likely to substantially underestimate the total number of trips as a result of the privacy algorithm that the operator has applied. The results described in the rest of this paper are therefore derived from analyses at either district (40 regions) or market area (12 regions) resolution.

Net travelling balance

To further validate the OD matrices, a comparison between the number of inbound and outbound trips was made for each district. It could be assumed that the net difference between these two measures during a full day ought to be close to zero (unless a trip spans over several 24-h periods).

This is confirmed in Fig. 5, which shows the cumulative inbound and outbound trips occurring during the analysed period for the 40 districts. The graphs show that the imbalances that occur during the day (e.g. as a result of people going to work) are rectified at the end of the day, implying that the number of inbound and outbound trip is close to equal.

In the 40 districts shown, the average gross traffic flow is $90{,}000 \pm 70{,}000$ passengers and the net traffic flow is 0 ± 1000, with 38 of the districts having total net deviations smaller than 1% of the respective traffic volume. The remaining two have deviations of -11% and $+13\%$. As such, the OD matrices are consistent with the assumption that the net traffic flow over a 24-h period is close to zero. Small deviations are to be expected, since not all travellers return to their origin at the end of a 24 h-period (for instance nightshift workers and travellers that spend the night elsewhere). Absolute numbers on the vertical axis are not shown to improve readability.

Validation against population statistics

Dataset A provides an hourly handset count per area. This count is fluctuating as people travel between different areas during the day. However, it could be assumed that the count should correspond to national population statistics at some point of the day (for instance, when all residents have returned home for the night). The boxplots shown in Fig. 6 show the relative hourly difference between population count estimated with the MPD and the official population statistics for the 40 districts. If all residents were indeed to spend a certain number of hours at their registered address each day, a substantial share of the observations ought to lay in the vicinity of the origin (i.e. entailing that the MPD count and population census concur). This, however, only occurs for a minority of the areas. For most of the districts, the number of residents exceeds the number of mobile signals, which can be explained by the fact that not all residents own mobile phones (for instance children). On the other hand, a few districts show opposite trends, notably that of St Hanshaugen. St Hanshaugen is a district located downtown, prominent with hotels (i.e. no official residents but many mobile signals) and popular with students, who tend to remain registered at their parents' place for financial reasons while living and studying in Oslo. Furthermore, many of the apartments there are rented out to e.g. foreign short-time workers (notably Swedes), who often remain registered in their home countries. This also entails that one apartment

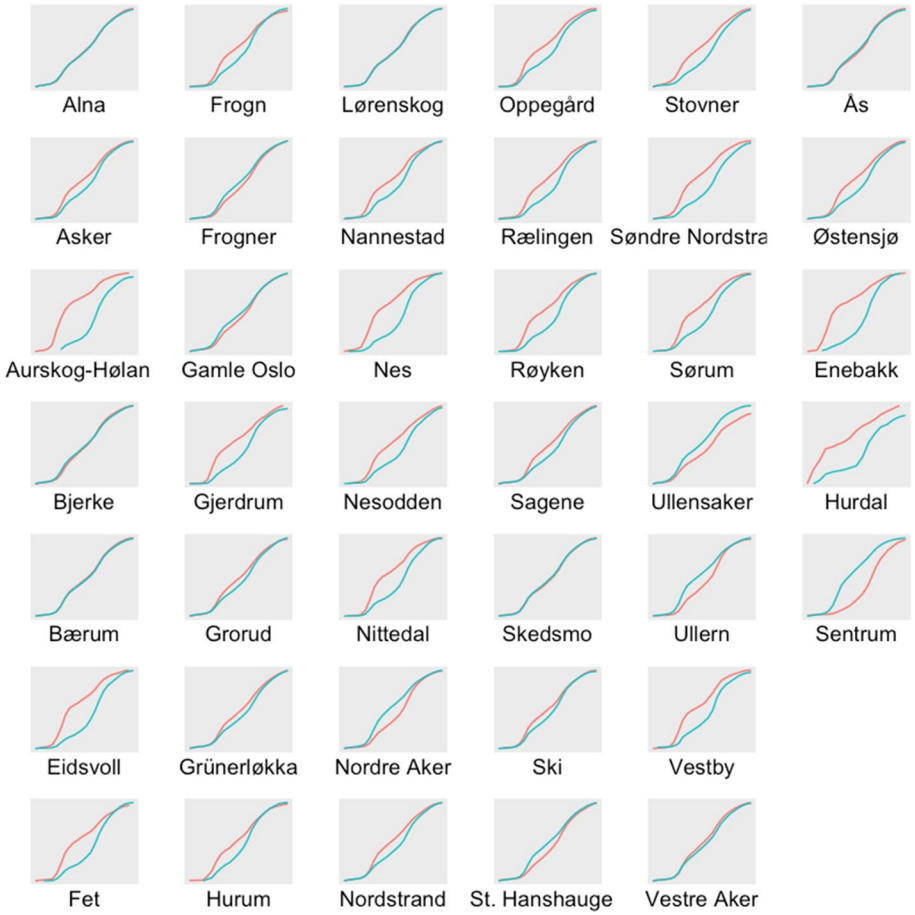
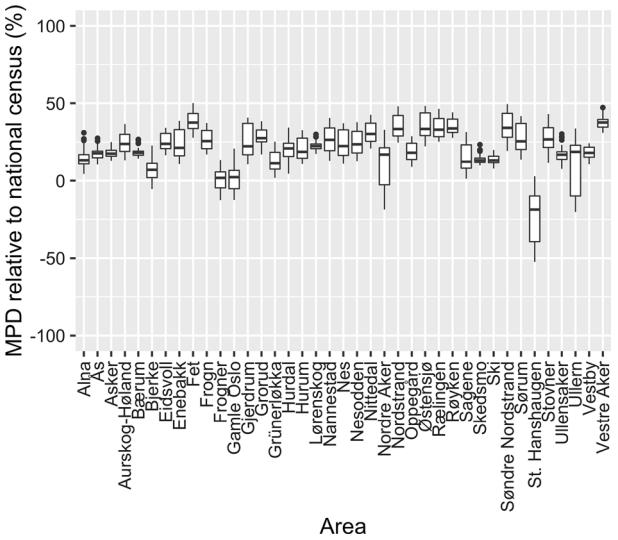


Fig. 5 Cumulative inbound and outbound trips for the 40 districts included in the study. The horizontal axis represents time (from 00:00 to 24:00) and the vertical axis shows the number of trips under taken (the scale is different in each graph)

may have one (or zero) official resident despite that it houses several inhabitants. Lastly, it is worth mentioning that national population statistics also carry inherent uncertainty; Statistics Norway estimate the accuracy of its census statistics to ± 5–10% (Schjalm 1996).

Validation against public transport data

To further validate the MPD, we compared the OD matrices with OD matrices constructed using door counts from public transportation, as described in the method section. Figure 7 shows the number of passengers travelling between the city centre of Oslo and the Frogner district during 1 day using both data sources. The shapes of the curves indicate that the distributions of the trips during the course of a day are similar. However, we do note that the centre-bound travel peak that occurs in the morning according to public transportation data does not appear in the MPD.



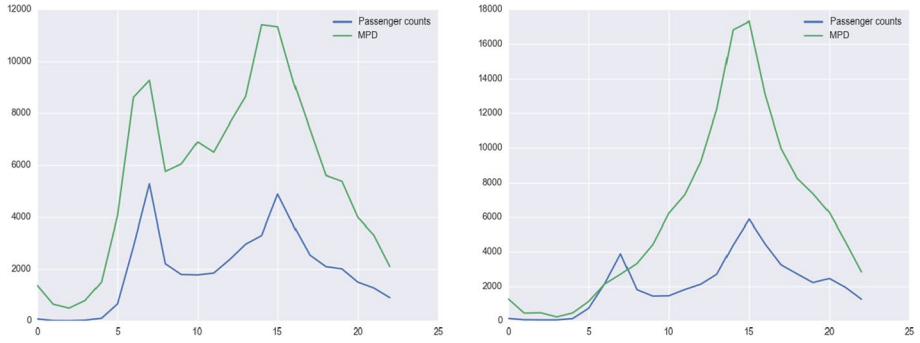


Fig. 6 Boxplot of hourly handset count per area as a percentage of the official population for the same area (census – MPD)/census * 100%

Fig. 7 Number of passengers travelling from the city centre of Oslo to the district of Frogner (left) and vice versa (right), during the course of a day, estimated using MPD and passenger counts respectively

We also find a discrepancy regarding the absolute number of trips, with the MPD leading to estimations two to three times higher than the estimations based on public transport. This is to some extent expected, as the MPD reports all undertaken trips while the latter only accounts for bus or tram trips. Figure 8 summarises the magnitude of the discrepancy between both metrics. Each dot represents a unidirectional trip between two subregions, and the values on the axis show the number of estimated trips obtained from passenger counts (horizontal axis) and MPD (vertical axis). As the study comprises 40 districts, this results in ${}^{40}\mathrm{P}_2 = \frac{40!}{(40-2)!} = 1560$ combinations/trips (where P symbolises the permutation operator). Most observations lie above the diagonal (blue line), indicating that OD matrices from MPD systematically lead to more observations.

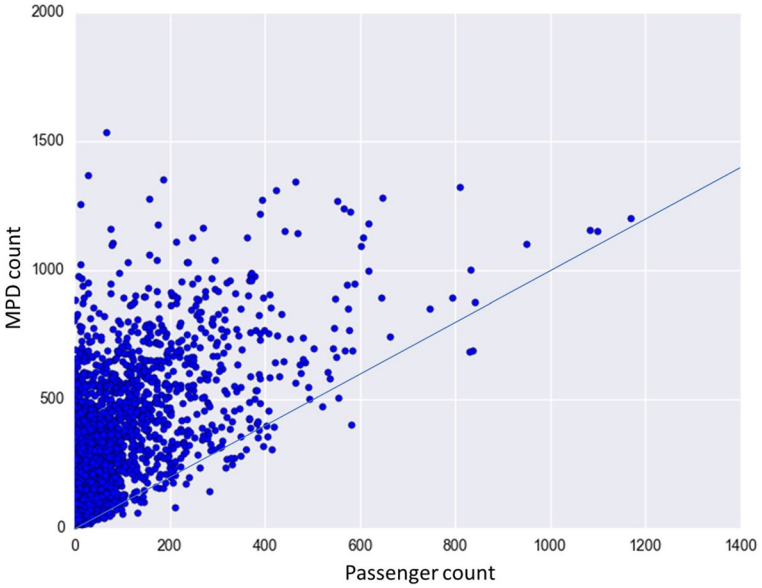


Fig. 8 Comparison between the travel counts estimated using MPD (vertical axis) and passenger counts (horizontal axis). Each dot represents a unidirectional trip between two districts

Validation against turnpike logs

The next validation procedure entails comparing the MPD to estimates of the number of eastbound travellers passing the highway turnpikes situated between the district of Bærum and the city centre based on the OD matrices stemming from MPD. The number was obtained by summing all trips originating from any of the western districts and ending in any of the other districts. This number was then compared with the vehicle count performed at the turnpikes (Fig. 9).

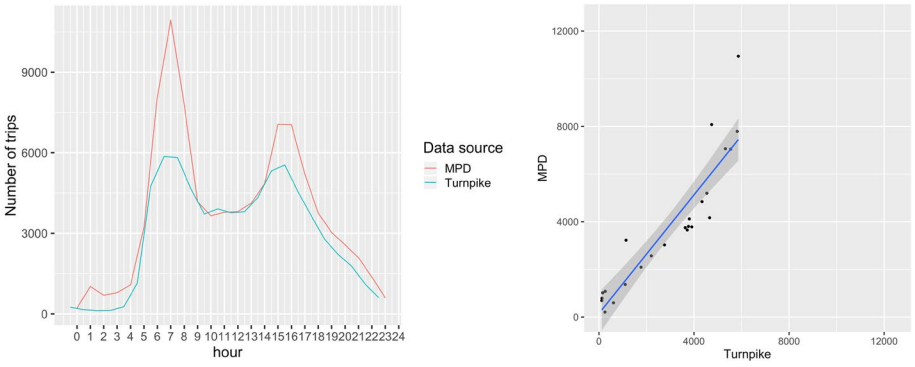


Fig. 9 Number of eastbound travellers passing the highway turnpike between the district of Bærum and the city centre (LHS graph), estimated with MPD data (red line) and based on vehicle counts (green line), and correlation plot of the two curves (RHS graph)

Akin to the comparison with the public transport data, the shapes of curves correlate well over the course of the day, but the total number of trips differ, with the MPD resulting in more trips. This, again, can be expected; the turnpike counter does not discern whether a vehicle contains one or several passengers, and not all trips are undertaken by car. For a short period, we see that MPD registers fewer trips than registered at the turnpikes. It is not clear from any of the two data sources and their metadata what causes this effect, but it is likely a consequence of temporal aggregation.

Validation against travel surveys

The final validation procedure consists in comparing the MPD-based OD matrices with OD matrices obtained from travel surveys. Such a comparison entails two important limitations. Firstly, the definition of a "trip" differs; while a travel survey respondent has perfect knowledge (in theory) of the distinct trips undertaken during a day, the OD matrices constructed from MPD require to mechanically define what constitutes a trip, the implications of which have already been discussed in the "Methods" section. Secondly, OD matrices do not account for trips occurring *within* one of the 40 districts of study (nor trips occurring outside the system boundaries of Oslo and Akershus), whereas survey respondents may account for such trips. As such, it is reasonable to expect that the total number of trips estimated from MPD lays below the estimations obtained from travel surveys.

The OD matrices provided by the operator resulted in 1.89 million trips for an average Wednesday (based on the four studied Wednesdays). With a population of around 1.33 million, this results in 1.4 trips per person per day. Alternatively, using the number of registered cell phones as a benchmark (1.08 million), the number still only amounts to 1.7 trips per person per day. According to a 2013/2014 report by Hjorthol et al. (2014), the average Norwegian citizen undertakes 3.3 trips per day, which confirms that the OD matrices significantly underestimate the number of reported trips, though it must be emphasised anew that these figures are based on different approaches and definitions and can therefore not be compared on the same basis.

To obtain an overall sense of how well the two OD sources correlate, we aggregated the 40 districts into the 12 market areas defined by the operator and plotted the undertaken trips in the matrices shown in Fig. 10, where values are normalised by origin so that the sum along the columns equal 100. Inter-market area trips were set to zero to enable comparison, and the district of West4 was excluded as it was not considered in the travel surveys. The plots are colour-scaled in shades of blue, with opaqueness decreasing with increasing shares.

Figure 10 indicates that the general distribution of trips seems to be consistent between the two data sources. In order to quantify the discrepancies, we constructed a distribution difference matrix (Fig. 11), which shows the cellwise difference between the two matrices from Fig. 10. The difference lies within a range of $[-13, 11]$ %, with more than half of the cells differing less than 1%. This indicates that the OD matrices obtained from MPD and travel surveys correlate well. We note that the largest differences occur for trips into the city centre, where MPD systematically underestimates the number of undertaken trips. This could be a result of the city centre area being relatively small (geographically), which increases the odds of misidentifying trips. We also note a general trend of larger discrepancies along the diagonal for regions that are geographically close (e.g. within the same cardinal directions, such as West3 to West2, Northeast4 to Northeast3, etc.). This entails that estimating OD matrices from MPD appears less

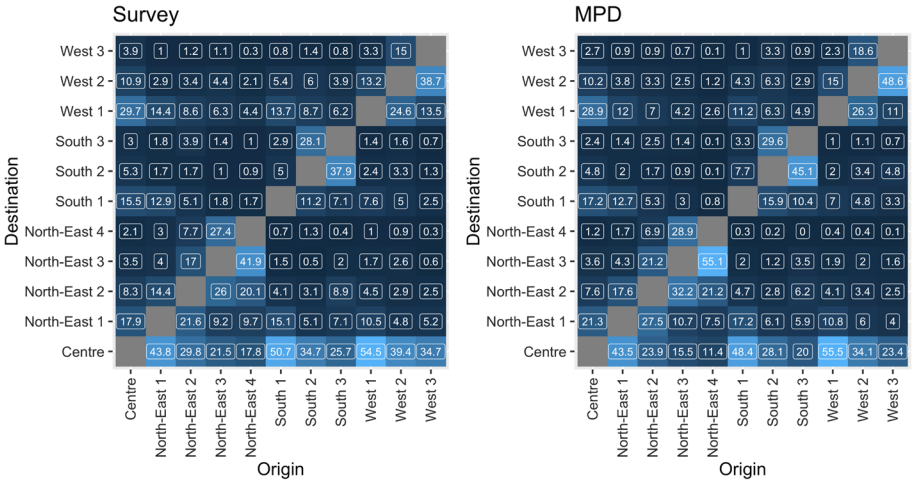
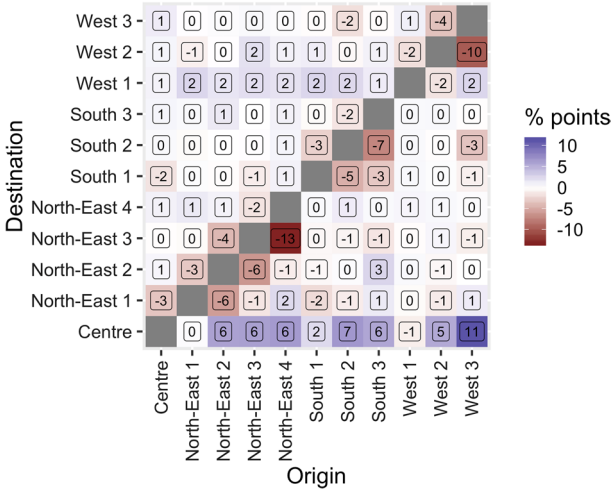


Fig. 10 Distribution of trips between different market areas (aggregated districts) using OD matrices from travel surveys (LHS graph) and MPD (RHS graph). Shares are normalised by origins (i.e. the sum along columns equal 100)

Fig. 11 Cellwise difference between the OD matrices displayed in this figure (difference in absolute percentage points)



precise for trips between neighbouring areas (under the assumption that OD matrices based on travel surveys constitute ground truth), although this is to be expected as a consequence of the low geographical precision.

A similar analysis was performed at the district level. Because the survey data was not as temporally precise as the OD matrices, the comparison could not be done in absolute numbers unless data was extrapolated. Hence, we performed the analysis on the same basis as for the market areas (shown in Fig. 10), and instead compared the share of trips originating from each origin to all destinations, for the MPD and the survey data respectively. Each point in Fig. 12 symbolises one OD combination (trip from district X to district Y). For instance, the uppermost point shows that of all trips originating from

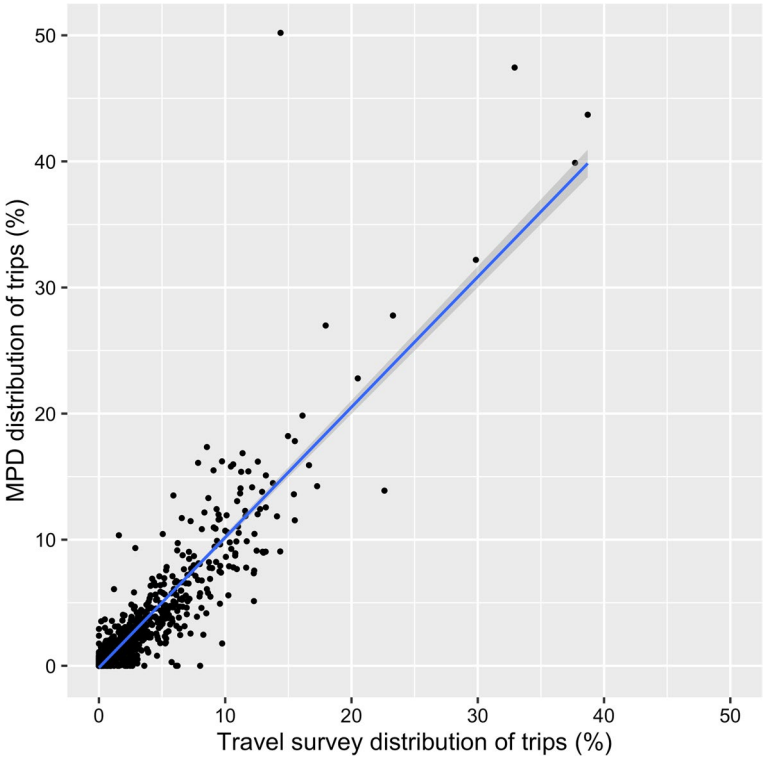


Fig. 12 Share of trips originating from each origin to all destinations, for trips compiled using MPD (vertical axis) and travel surveys (horizontal axis). Each point symbolises one unique OD (e.g. district X to district Y), and the percentage is the share of all trips starting in district X that end in district Y

district X, 50% end in district Y according to the MPD, but only 14% according to the survey data. While some distinct outliers can be observed, the distribution of trips from the two sources correlate well, with a coefficient of determination (R^2) of 0.82.

Discussion

MPD constitutes a new and vast source of information that entails new opportunities for the fields of transport network designs and public infrastructure planning. The unprecedented size and comprehensiveness of MPD compared to traditional sources used to estimate OD matrices make it an invaluable complement to e.g. travel surveys, which typically contain only a few thousand samples, and to travel card logs, which only cover travellers using a particular transport mode. As a result, MPD has been praised in the literature as a virtual gold mine for transport planning and modelling. While the sheer amount of data may be unprecedented, its usefulness in the estimation of reliable OD matrices is still unclear. The various problems associated with the use of MPD in the estimation of OD matrices have already been thoroughly described in the introduction of this paper, and while their extent depends largely on the characteristics of the studied area and the quality of the MPD used, the inherent nature of MPD entails a range of systematic issues that inevitably hamper its

robustness. Notably, coarseness of data and privacy concerns restrict the resolution to relatively large geographical areas. In this study, we have addressed the latter characteristic by assessing the robustness of a set of OD matrices that had been pre-compiled (including anonymised and censored) by the mobile phone operator prior to be distributed for analysis. We compared the data to several other sources, including travel surveys.

When studying the OD matrices at borough level, we found that a substantial share of the trips had been censored by the privacy algorithm applied by the mobile phone operator, rendering those matrices incomplete and unreliable for describing travel patterns. Furthermore, our results reflected the complexities involved in defining what constitutes a trip; a trip defined by a travel survey (e.g. an individual travelling from A to B) does not necessarily coincide with a trip as defined by MPD, especially for relatively short trips involving stopovers. This was confirmed in our results, where discrepancies between OD matrices estimated from travel surveys and MPD-based OD matrices tended to be larger for trips between neighbouring areas (despite choosing a relatively coarse resolution). This concurs with findings from other studies as well; for instance, Alexander et al. (2015) find good correlation of MPD with travel surveys when aggregating trip origins and destinations to areas larger than one square mile.

Perhaps the greatest challenge in assessing the reliability of MPD is the lack of exact ground truth. As seen in this study and others (e.g. Mamei et al. (2019)), matching MPD to other sources entails multiple challenges. All potential validation sources suffer from inherent issues of incompleteness and/or inaccuracy; public transport logs cover only passengers on buses and trams and often do not provide enough detail to accurately describe the undertaken trips (e.g. because cards are only validated on entry, because of monthly subscriptions, etc.); traffic cameras/sensors capture only vehicle passengers and cannot determine the amount of passengers in each vehicle; travel surveys involve selection bias and misreporting, etc. Comparing MPD against other sources linked to specific transport modes always entails a comparison between sample data of different sizes (as more people own mobile phones than e.g. travel by car), and comparing distribution curves as we have done in this study implicitly assumes that the percentage of phone owners (and number of phones per user) in each group (e.g. car drivers and bus passengers) is equal. These relatively large and variable errors, combined with a lack of quality metrics for MPD-sourced trip generation (Huang et al. 2019), makes direct comparison of algorithms and models difficult.

Additionally, practical challenges regarding use of MPD in transport planning remain. MPD is currently not freely available in most countries and needs to be purchased from mobile operators. This is usually costly, which has resorted many published studies to purchase data from specific periods and specific areas (typically a city and/or the surrounding agglomeration) in order to answer specific research questions. Constructing a dynamic transport model based on regularly updated travel information would require continuous (or at least semi-continuous) access to MPD and would therefore entail a significant cost. However, this might change in the future if the use of MPD for transport-related purposes increases, potentially creating a well-functioning market for storage, exchange, and pricing of relevant products.

In this study, we only considered data and transformation methods that were commercially available, and only stemming from one mobile phone operator. These data sources lag behind some of the research frontier, for instance with respect to transport mode detection (Huang et al. 2019), which are currently not available for purchase. While this is of high importance for transport planning, the results from such advanced methods emerging from the research field seem to be more on the level of "proof of concept" rather than

"proof of value" or business ready. As such, we have relied on the less complex approaches of trip identification and population scaling, limited by k-anonymity. Nevertheless, we found that the lack of coherent and transparent definitions and approaches between mobile phone operators makes the work of establishing quality metrics more important.

Concluding remarks and topics for further research

In this paper, we have studied a set of pre-compiled OD matrices of the greater Oslo area estimated using MPD and validated them against several other sources including passenger counts from public transportation, population census and highway traffic logs. We also compared the matrices with OD matrices constructed using individual travel surveys to assess whether such pre-compiled matrices could be used as a potential replacement for travel surveys, which have traditionally been the source of choice in the estimation of OD matrices.

We found that the scaled-up results from the OD matrices stemming from MPD and from travel surveys concur well in terms of travel patterns, but that the number of undertaken trips differ substantially. We also found that the accuracy of the used MPD (in terms of how well it correlates with travel surveys) was lower for trips between neighbouring areas, differing up to 13% even when a relatively coarse resolution was chosen. The accuracy was also shown to be lower for trips bound for the city centre of Oslo as a result of that area being geographically small, which increases the odds of misidentifying trips. We therefore concluded that such as set of pre-compiled MPD-based OD matrices could potentially be useful for mapping long-distance trips, but that additional data would be needed to accurately identify shorter trips.

The results of this study show that while MPD does constitute a vast and worthwhile resource for transport-related applications, the usefulness rapidly decreases at finer resolutions when privacy regulations limit the data available to transport researchers. In the case where OD matrices have been pre-compiled by the operator and where a relatively stringent privacy algorithm has been used to filter out data, we concluded that MPD does not constitute a comprehensive alternative to travel surveys for transport operators but should rather be considered a complementary source of information. We therefore suggest several areas of potential future research that could contribute to the further advancement and improvement of MPD-based transport-related analyses. Firstly, methods should be developed to harvest the full potential of MPD while still respecting governing privacy regulations. Cut-off thresholds may be suitable for large study areas and/or longer periods of time, but other anonymisation schemes should be developed for fine-grained studies, for instance based on those suggested by Machanavajjhala et al. (2008) and Chatzikokolakis et al. (2017). Secondly, the robustness and validity of MPD should be compared to ground truth data in more detail. This will entail several varied, controlled experiments, where detailed information about the transport activity is collected simultaneously with MPD observations. Thirdly, business models for MPD data transactions and standardisation for MPD products should be developed, for the benefit of both MPD suppliers and MPD users, as the current complex pricing scheme on MPD and the lack of a commonly accepted standard for MPD format, content, and quality constitutes a fundamental impediment to transport research based on MPD.

Author contributions Journal submission guidelines do not indicate that these are mandatory.

Funding No funding to declare.

Availability of data and material Data and material used have not been made available due to privacy requirements.

Code availability Code used has not been made available.

Compliance with ethical standards

Conflict of interest The authors declare no conflict of interest.

References

Aguiléra, V., Allio, S., Benezech, V., Combes, F., Milion, C.: Using cell phone data to measure quality of service and passenger flows of Paris transit system. Transp. Res. Part C Emerg. Technol. **43**, 198–211 (2014)

Ahas, R., Aasa, A., Silm, S., Tiru, M.: Daily rhythms of suburban commuters' movements in the Tallinn metropolitan area: case study with mobile positioning data. Transp. Res. Part C Emerg. Technol. **18**, 45–54 (2010)

Aksehirli, E., Li, Y.: Predicting MRT trips in Singapore by creating a mobility behavior model based on GSM data. In: 2018 IEEE International Conference on Data Mining Workshops (ICDMW), pp. 632–639. IEEE (2018)

Alexander, L., Jiang, S., Murga, M., González, M.C.: Origin–destination trips by purpose and time of day inferred from mobile phone data. Transp. Res. Part C Emerg. Technol. **58**, 240–250 (2015)

Asgari, F.: Inferring User Multimodal Trajectories from Cellular Network Metadata in Metropolitan Areas. Institut National des Télécommunications, Évry (2016)

Bachir, D., Khodabandelou, G., Gauthier, V., El Yacoubi, M., Puchinger, J.: Inferring dynamic origin–destination flows by transport mode using mobile phone data. Transp. Res. Part C Emerg. Technol. **101**, 254–275 (2019)

Bassolas, A., Ramasco, J.J., Herranz, R., Cantú-Ros, O.G.: Mobile phone records to feed activity-based travel demand models: MATSim for studying a cordon toll policy in Barcelona. Transp. Res. Part A Policy Pract. **121**, 56–74 (2019)

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

Calabrese, F., Di Lorenzo, G., Liu, L., Ratti, C.: Estimating Origin–Destination Flows Using Opportunistically Collected Mobile Phone Location Data from One Million Users in Boston Metropolitan Area (2011a)

Calabrese, F., Di Lorenzo, G., Liu, L., Ratti, C.: Estimating origin–destination flows using mobile phone location data. IEEE Pervasive Comput. **10**, 36–44 (2011b)

Calabrese, F., Diao, M., Di Lorenzo, G., Ferreira Jr., J., Ratti, C.: Understanding individual mobility patterns from urban sensing data: a mobile phone trace example. Transp. Res. Part C Emerg. Technol. **26**, 301–313 (2013)

Chatzikokolakis, K., Elsalamouny, E., Palamidessi, C., Pazii, A.: Methods for Location Privacy: A comparative overview. In: Foundations and Trends® in Privacy and Security, vol. 1, no. 4, pp. 199–257. Now publishers inc. https://doi.org/10.1561/3300000017 (2017)

Chen, C., Ma, J., Susilo, Y., Liu, Y., Wang, M.: The promises of big data and small data for travel behavior (aka human mobility) analysis. Transp. Res. Part C Emerg. Technol. **68**, 285–299 (2016)

Danafar, S., Piorkowski, M., Krysczcuk, K.: Bayesian framework for mobility pattern discovery using mobile network events. In: 2017 25th European Signal Processing Conference (EUSIPCO), pp. 1070–1074. IEEE (2017)

Di Lorenzo, G., Sbodio, M., Calabrese, F., Berlingerio, M., Pinelli, F., Nair, R.: Allaboard: visual exploration of cellphone mobility data to optimise public transport. IEEE Trans. Visual Comput. Graph. **22**, 1036–1050 (2015)

Doyle, J., Hung, P., Kelly, D., Mcloone, S.F., Farrell, R.: Utilising Mobile Phone Billing Records for Travel Mode Discovery (2011)

Drageide, V.: Towards Privacy Management of Information Systems. The University of Bergen, Bergen (2009)
Forrest, T.L., Pearson, D.F.: Comparison of trip determination methods in household travel surveys enhanced by a global positioning system. Transp. Res. Rec. **1917**, 63–71 (2005)
García, P., Herranz, R., Javier, J.: Big data analytics for a passenger-centric air traffic management system. Presented at the 6th SESAR Innovation Days, Delft, Netherlands (2016)
García-Albertos, P., Picornell, M., Salas-Olmedo, M.H., Gutiérrez, J.: Exploring the potential of mobile phone records and online route planners for dynamic accessibility analysis. Transp. Res. Part A Policy Pract. **125**, 294–307 (2019)
Gundlegård, D., Rydergren, C., Breyer, N., Rajna, B.: Travel demand estimation and network assignment based on cellular network data. Comput. Commun. **95**, 29–42 (2016)
Hjorthol, R., Engebretsen, Ø., Uteng, T.P.: Den nasjonale Reisevaneundersøkelsen 2013/14: Nøkkelrapport. Transportøkonomisk institutt, Oslo (2014)
Holleczek, T., Yin, S., Jin, Y., Antonatos, S., Goh, H.L., Low, S., Shi-Nash, A.: Traffic measurement and route recommendation system for mass rapid transit (MRT). In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1859–1868. ACM (2015)
Horn, C., Kern, R.: Deriving public transportation timetables with large-scale cell phone data. Procedia Comput. Sci. **52**, 67–74 (2015)
Huang, Z., Ling, X., Wang, P., Zhang, F., Mao, Y., Lin, T., Wang, F.-Y.: Modeling real-time human mobility based on mobile phone and transportation data fusion. Transp. Res. Part C Emerg. Technol. **96**, 251–269 (2018)
Huang, H., Cheng, Y., Weibel, R.: Transport mode detection based on mobile phone network data: a systematic review. Transp. Res. Part C Emerg. Technol. **101**, 297–312 (2019)
Iqbal, M.S., Choudhury, C.F., Wang, P., González, M.C.: Development of origin–destination matrices using mobile phone call data. Transp. Res. Part C Emerg. Technol. **40**, 63–74 (2014)
Isaacman, S., Becker, R., Cáceres, R., Kobourov, S., Martonosi, M., Rowland, J., Varshavsky, A.: Ranges of human mobility in Los Angeles and New York. In 2011 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops), pp. 88–93. IEEE (2011)
Kalatian, A., Shafahi, Y.: Travel mode detection exploiting cellular network data. In: MATEC Web of Conferences, pp. 03008. EDP Sciences (2016)
Larijani, A.N., Olteanu-Raimond, A.-M., Perret, J., Brédif, M., Ziemlicki, C.: Investigating the mobile phone data to estimate the origin destination flow and analysis; case study: Paris region. Transp. Res. Procedia **6**, 64–78 (2015)
Li, G., Chen, C.-J., Peng, W.-C., YI, C.-W.: Estimating crowd flow and crowd density from cellular data for mass rapid transit. In: Proceedings of the 6th International Workshop on Urban Computing (in Conjunction with ACM KDD 2017) (2017)
Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J., Vilhuber, L.: Privacy: theory meets practice on the map. In: 2008 IEEE 24th International Conference on Data Engineering, pp. 277–286. IEEE (2008)
Mamei, M., Bicocchi, N., Lippi, M., Mariani, S., Zambonelli, F.: Evaluating origin–destination matrices obtained from CDR data. Sensors (Basel) **19**, 4470 (2019)
Montero, L., Ros-Roca, X., Herranz, R., Barceló, J.: Fusing mobile phone data with other data sources to generate input OD matrices for transport models. Transp. Res. Procedia **37**, 417–424 (2019)
Ni, L., Wang, X.C., Chen, X.M.: A spatial econometric model for travel flow analysis and real-world applications with massive mobile phone data. Transp. Res. Part C Emerg. Technol. **86**, 510–526 (2018)
Phithakkitnukoon, S., Sukhvibul, T., Demissie, M., Smoreda, Z., Natwichai, J., Bento, C.: Inferring social influence in transport mode choice using mobile phone data. EPJ Data Sci. **6**, 11 (2017)
Poonawala, H., Kolar, V., Blandin, S., Wynter, L., Sahu, S.: Singapore in motion: Insights on public transport service level through farecard and mobile data analytics. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and data mining, pp. 589–598. ACM (2016)
Qu, Y., Gong, H., Wang, P.: Transportation mode split with mobile phone data. In: 2015 IEEE 18th International Conference on Intelligent Transportation Systems, pp. 285–289. IEEE (2015)
Schjalm, A.: Kvalitetsundersøkelsen for Folke- og boligtelling 1990. In: Norway, S. (ed.) Oslo—Kongsvinger: Statistics Norway (1996)
Schlaich, J., Otterstätter, T., Friedrich, M.: Generating trajectories from mobile phone data. In: Proceedings of the 89th Annual Meeting Compendium of Papers. Transportation Research Board of the National Academies (2010)
Smoreda, Z., Olteanu-Raimond, A.-M., Couronné, T.: Spatiotemporal data from mobile phones for personal mobility assessment. Transp. Surv. Methods Best Pract. Decis. Mak. **41**, 745–767 (2013)

Sørensen, A.Ø., Bjelland, J., Bull-Berg, H., Landmark, A.D., Akhtar, M.M., Olsson, N.O.: Use of mobile phone data for analysis of number of train travellers. J. Rail Transp. Plan. Manag. **8**, 123–144 (2018)
Statistics Norway: Classification of Statistical Tract and Basic Statistical Unit (2019). Available: https://www.ssb.no/en/klass/klassifikasjoner/1. Accessed 29 Nov 2019
Stopher, P.R., Greaves, S.P.: Household travel surveys: Where are we going? Transp. Res. Part A Policy Pract. **41**, 367–381 (2007)
Stopher, P., Fitzgerald, C., Xu, M.: Assessing the accuracy of the Sydney Household Travel Survey with GPS. Transportation **34**, 723–741 (2007)
Toole, J.L., Colak, S., Sturt, B., Alexander, L.P., Evsukoff, A., González, M.C.: The path most traveled: travel demand estimation using big data resources. Transp. Res. Part C Emerg. Technol. **58**, 162–177 (2015)
Vazifeh, M.M., Zhang, H., Santi, P., Ratti, C.: Optimizing the deployment of electric vehicle charging stations using pervasive mobility data. Transp. Res. Part A Policy Pract. **121**, 75–91 (2019)
Wang, F., Chen, C.: On data processing required to derive mobility patterns from passively-generated mobile phone data. Transp. Res. Part C Emerg. Technol. **87**, 58–74 (2018)
Wang, H., Calabrese, F., Di Lorenzo, G., Ratti, C.: Transportation mode inference from anonymized and aggregated mobile phone call detail records. In: 13th International IEEE Conference on Intelligent Transportation Systems, pp. 318–323. IEEE (2010)
Wang, Z., He, S.Y., Leung, Y.: Applying mobile phone data to travel behaviour research: a literature review. Travel Behav. Soc. **11**, 141–155 (2018)
Wolf, J., Loechl, M., Thompson, M., Arce, C.: Trip rate analysis in GPS-enhanced personal travel surveys. In: Stopher, P.R., Jones, P. (eds.) Transport Survey Quality and Innovation. Emerald Group Publishing Limited, Bingley (2003)
World Bank: Mobile Cellular Subscriptions (per 100 People) (2019). Available: https://data.worldbank.org/indicator/it.cel.sets.p2. Accessed 29 Nov 2019
Wu, W., Cheu, E.Y., Feng, Y., Le, D.N., Yap, G.E., Li, X.: Studying intercity travels and traffic using cellular network data. In: Mobile Phone Data for Development: Net Mob 2013 (2013)
Wu, C., Thai, J., Yadlowsky, S., Pozdnoukhov, A., Bayen, A.: Cellpath: fusion of cellular and traffic sensor data for route flow estimation via convex optimization. Transp. Res. Part C Emerg. Technol. **59**, 111–128 (2015)
Wu, L., Yang, B., Jing, P.: Travel mode detection based on GPS raw data collected by smartphones: a systematic review of the existing methodologies. Information **7**, 67 (2016)
Xu, C., Ji, M., Chen, W., Zhang, Z.: Identifying travel mode from GPS trajectories through fuzzy pattern recognition. In: 2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery, pp. 889–893. IEEE (2010)
Yamada, Y., Uchiyama, A., Hiromori, A., Yamaguchi, H., Higashino, T.: Travel estimation using control signal records in cellular networks and geographical information. In: 2016 9th IFIP Wireless and Mobile Networking Conference (WMNC), pp. 138–144. IEEE (2016)
Zheng, Y., Chen, Y., Li, Q., Xie, X., Ma, W.-Y.: Understanding transportation modes based on GPS data for web applications. ACM Trans. Web (TWEB) **4**, 1 (2010)

Andreas Dypvik Landmark is a Senior research scientist in SINTEF Digital. His background is in computer science from the Norwegian University of Science and Technology. He has worked for the last decade on transportation using novel data sources in railway. Primary research interest is in data-driven decision support for operational management of transport systems.

Petter Arnesen holds a PhD in statistics and M.Sc. in industrial mathematics from the Norwegian University of Science and Technology. He is a senior researcher at SINTEF working with analysis of data from the transport sector, including sensor data and travel behaviour, and in particular working towards the field of intelligent transport systems.

Carl-Johan Södersten has a M.Sc. in Engineering Mathematics from Chalmers University of Technology and a PhD in Industrial Ecology from the Norwegian University of Science and Technology. He is currently working as a research scientist in the field of intelligent transport systems, specialising in data processing, modelling and analysis.

Odd André Hjelkrem received his PhD in Transportation Engineering (2016) and M.Sc. in Applied Physics (2007) from the Norwegian University of Science and Technology, and has worked as a researcher at SINTEF since 2007. His research interests are vehicle technology, decarbonization of the mobility sector and mathematical modelling of transport demand and energy use.