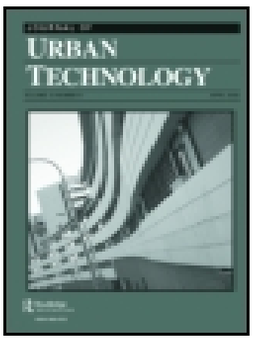

Routledge
Taylor & Francis Group

Journal of Urban Technology

ISSN: 1063-0732 (Print) 1466-1853 (Online) Journal homepage: http://www.tandfonline.com/loi/cjut20

Improving A Priori Demand Estimates Transport Models using Mobile Phone Data: A Rotterdam-Region Case

L.J.J. Wismans, K. Friso, J. Rijsdijk, S.W. de Graaf & J. Keij

To cite this article: L.J.J. Wismans, K. Friso, J. Rijsdijk, S.W. de Graaf & J. Keij (2018): Improving A Priori Demand Estimates Transport Models using Mobile Phone Data: A Rotterdam-Region Case, Journal of Urban Technology, DOI: 10.1080/10630732.2018.1442075

To link to this article: https://doi.org/10.1080/10630732.2018.1442075

Published online: 06 Apr 2018.

JOURNAL OF URBAN TECHNOLOGY, 2018
https://doi.org/10.1080/10630732.2018.1442075

Improving A Priori Demand Estimates Transport Models using Mobile Phone Data: A Rotterdam-Region Case

L.J.J. Wismans[a], K. Friso[b], J. Rijsdijk[c], S.W. de Graaf[d], and J. Keij[e]

[a]Centre for Transport Studies, University of Twente, Deventer, Netherlands; [b]DAT.Mobility, Deventer, Netherlands; [c]Municipality of Rotterdam, Rotterdam, Netherlands; [d]Goudappel Coffeng, Deventer, Netherlands; [e]Mezuro, Weesp, Netherlands

ABSTRACT

Mobile phone data are a rich source to infer all kinds of mobility-related information. In this research, we present an approach where mobile phone data are used and analyzed for enriching the transport model of the region of Rotterdam. In this research Call Detail Records (CDR) are used from a mobile phone provider in the Netherlands that serves between 30 and 40 percent of Dutch mobile phones. Accessing these data provides travel information of about one-third of the Dutch population. No other data source is known that gives travel information at a national scale at this high level. The raw data of one month is processed into basic information which is subsequently translated into OD-information (Origin-Destination) based on several decision rules. This OD information is compared with the traditionally estimated a priori OD matrix of the Rotterdam transport model and the Dutch yearly national household travel survey. Based on the analysis and assignment results, an approach is developed to combine the mobile phone OD-information and an a priori OD matrix using the best of both worlds. Results show a better match of the assignment results of this matrix with the counts indicating a better quality of the matrix.

KEYWORDS

Transport modelling; mobile phone data; demand modelling

Introduction

General

Detailed data of individual activities and interactions are currently being collected at an unprecedented spatial and temporal granularity level ranging from data collected by mobile phones to social media services. These big datasets have a high potential value for monitoring, planning, and managing transport systems; however, they are still not too often used in the transport field. Today, transport planners and engineers rely heavily on transport demand models for their understanding of travel behavior and the effectiveness of infrastructure investments. These models provide quantitative

CONTACT L.J.J. Wismans l.j.j.wismans@utwente.nl

information on current and future travel patterns (e.g., destination and mode choice) and traffic conditions in peak hours (Bliemer et al., 2014; Van Eck et al., 2014). These models use the same type of data and in the same way as were used decades ago (Benbow et al., 2008, Joksimovic and Hofman, 2014). The main basis for these models are census data and travel surveys describing one weekday for each respondent in "representative" periods when traffic flows are maximal (Ortúzar et al., 2011). In the state of practice an a priori OD matrix is constructed based on these census and household data and then in a next step, they are further calibrated against traffic counts (often based on loop detector data) or possible other available measurements within the network to estimate the final a posteriori OD-matrix following a (multi-objective) optimization procedure. Collecting these travel surveys is a time-consuming task and they only provide a snapshot of travel behavior, because they cover a relatively small sample of all trips in terms of the number of participants as well as the number of days covered. As a result, these kinds of surveys are only repeated after several years and often only affordable by authorities. In most cases the travel surveys also underestimate the number of trips and show a bias in the types of trips being reported because people are asked to report the trips themselves. The research of Stopher and Greaves (2007) shows, for example, that household travel surveys omit a significant number of trips (between 20 and 30 percent depending on the method of retrieval). Furthermore, the gravity model often used in the distribution phase of constructing the a priori matrix assumes that trips are being made depending on size of production and attraction and generalized costs between zones (i.e., maximization of entropy). As a result, this method cannot correctly estimate the number of trips being made between zones that are traditionally more connected. Calibration of this a priori matrix against traffic counts is essential in estimating the final a posteriori OD matrix. This calibration procedure can solely focus on minimizing the difference between assignment results of the a posteriori OD-matrix and counts using the a priori OD-matrix as a starting point or a multi-objective optimization procedure minimizing the difference between the a priori and a posteriori OD-matrix as well. Therefore, the quality of the a priori matrix also determines the quality of the transport model. If the calibration impact is large (i.e., the difference between the a priori and a posteriori matrix) minimizing differences with counts, the estimations of the first three steps of the transport model are questionable which also reflects on their use for predictions of future states. The structure of this a priori OD matrix (i.e., relations between the zones) is therefore crucial, because in both cases, the calibration procedure does not necessarily improve this structure. As a result, the calibration procedure does not correct for possible biases in the survey data nor the possible errors made in the distribution phase. Introducing the use of mobile phone data to improve the estimation of the a priori OD matrix can improve the structure of the matrix and, therefore, the quality of the a posteriori OD matrix and predictions of the transport model as well.

The Potential of Using Mobile Phone Data

There is a big potential in using mobile phone data to improve transport models, because they can provide much larger samples and offer almost continuous measurements (24/7/365), as is also confirmed by Von Morner (2017). In this research, we will focus on Call Detail Records (CDR), because these mobile phone data are stored by mobile phone

operators for billing purposes and can, therefore, be used everywhere if made available. CDR data contain event-based information on the communication of a specific device (call, text and in most cases also data) with a cell tower including the ID of the cell tower, which offers the opportunity to connect a device with a location. Analyzing the time series of CDR data of a device offers the opportunity to extract its movement and determine the trips being made. Furthermore, it offers the opportunity to better understand the day-to-day dynamics and updating transport models more frequently, providing more accurate decision information, making it possible to address problems faster, but also addressing problems of which little was known (e.g., non-regular traffic conditions) and addressing problems that are difficult to model. Because collecting these data is less time consuming and costly, it also offers the opportunity to use this type of information for other services provided by private companies (e.g., for marketing purposes).

Earlier Research

Mobile phone data might be used for deriving various transport-related indicators (e.g., number of trips, mode choice, and travel times). Bar-Gera (2007) compared, for example, the use of mobile phone data to determine traffic speeds and travel time with speeds and travel times derived from loop detector data. The performance was shown to be adequate, although relative differences were found of more than 20 percent. Furthermore, the author also had access to (silent) handover data, which is not generally available in CDR data. Caceras et al. (2007) discusses several other projects and research in which mobile phone data are used. In a few cases, the validity of the outcome was researched. In these cases, the derivation of traffic operational indicators from mobile phone data shows better results when traffic was monitored on long stretches (e.g., travel times on corridors) as a result of the location error associated with mobile phone data. This indicates that deriving Origin-Destination (OD) information monitoring trip movements from mobile phone data for transport modelling data is of high potential.

Von Morner (2017) indicates that there is a considerable body of research done on the generation of OD matrices based on mobile phone data, showing promising results in this area. Already in 2002, White and Wells (2002) published research in which the authors suggested that it is possible to extract an OD matrix from mobile phone data, but were still in a research phase. The potential was indicated, but it was concluded that more research was needed to actually derive such an OD matrix. Caceras et al. (2007) simulated mobile phone data and used these data to construct an OD matrix. They also showed the high potential of using these types of data for this purpose with results that are more cost-effective than traditional data collection methods using the mobile phone infrastructure which is already in place. Although the simulation of mobile phone data offered the opportunity to validate their methods (i.e., the ground truth is known), the question is whether the authors could simulate all possible flaws of the system connected. Moreover, like Bar-Gera (2007) the authors assumed that silent handovers were also known. Calabrese et al. (2011) was one of the first to produce an OD matrix from an actual detailed mobile phone data set for the Boston Region, showing that the estimated OD flows correlate well with census data. Ma et al. (2013) estimated an OD matrix combining the trips derived from mobile phone data with calibration techniques using count data. Although they showed accurate results in representing counts with this approach, the use of the

same count data to calibrate the matrix does influence the actual value of this validation approach. Research of Nanni et al. (2013) also showed that it is possible to create an OD matrix with real mobile phone data in a country with hardly reliable statistics (i.e., Ivory Coast) showing mobility patterns, both nationally and within the capital Abijan. Gundlegard et al. (2016) used the same dataset as Nanni, provided as part of the D4D challenge to construct an OD matrix and development of mobility metrics. They also explain and provide a nice overview of challenges in deriving travel demand and routes based on mobile phone data. Obviously also in this case no validation of their methods was possible, because of the lack of data. In similar research for the Senegal network, optimizations were proposed based on mobility patterns derived from mobile phone data (De Romph et al., 2015). However, because of the lack of data in this case as well, the accuracy of the resulting patterns was not evaluated. Iqbal et al. (2014) developed a method for deriving an OD matrix using mobile phone data and applied their method to Dhaka, Bangladesh. They investigated the validity of their estimated OD matrix by assigning this matrix using a microscopic simulation, comparing the assignment results with traffic counts. However, because counts were also used to calibrate the model (also to increase trips derived from mobile phone data to absolute levels), the validation results could have been highly influenced by the calibration process. Huntsinger and Ward (2015) developed an external trip model using mobile phone data showing promising results compared with household travel surveys. Çolak et al. (2015) presented a data treatment pipeline that uses mobile phone data and population density to generate trip matrices in two metropolitan areas (i.e., Boston and Rio de Janeiro) showing comparable results with existing information reported in local travel surveys in Boston and existing origin destination matrices in Rio de Janeiro. However, this comparison focused on a highly aggregate level and trip length distribution or possible biases in OD information derived from mobile phone data was not considered. Bonnel et al. (2015) compared an OD matrix derived from mobile phone data with traditional census data and household travel surveys for the Paris Region. The authors address the lack of validation of earlier research attempting to derive an OD matrix using mobile phone data. In their research they tested the potential and showed high correlations between the household surveys and the OD matrix derived from mobile phone data. However, they also pointed out that on specific OD relations the differences can be very large. Alexander et al. (2015) developed a method to derive OD matrices by purpose and time of day using mobile phone data. They found that when aggregating origins and destination on town scale, trip estimates show a higher correlation (over 0.95) with survey data resulting in good average daily activity and trip representation, than home Census area definitions of Boston (on average 0.50) and recommended evaluation of assigning OD matrices derived from mobile phone data to explore how to improve existing transport planning models.

The use of mobile phone data introduces possible biases and errors. Calabrese et al. (2011) stress that the localization error limits the minimum size of the regions which can be considered. As a result, there is a possible mismatch in the level of detail in zone definition needed in the transport model versus the zoning possible with mobile phone data, which needs to be addressed. Furthermore, the market share of the mobile phone operator in combination with people not owning a mobile phone, the potential non-randomness of the mobile phone users in combination with the fact that the CDR are event-driven and the number of devices a person carries are possibly causing a bias within the

mobile phone dataset. Although these biases and errors are acknowledged in earlier research and recently stressed in the research of Zhao et al. (2016), they are often not addressed or corrected. Bonnel et al. (2015), for example, made the strong assumption that they compensate each other, because these biases are to some extent contradictory. Furthermore, Tolouei et al. (2017) shows that an OD matrix derived from mobile phone data when adjusted using independent data sources to address various known limitations and biases of this data, does not seem to be either biased or less accurate than traditional OD matrix estimation methods, as well as that OD matrices derived from mobile phone data could result in a more consistent estimate of trips for those areas where no road side interviews are available. In this research the OD matrices derived from mobile phone data were also calibrated using counts before validating the results.

This Research

In all previous research, mobile phone data were used to extract an OD matrix, showing promising results. In many cases no validation had been done and if done simulated data were used or a comparison was made with census or household survey data or traffic counts were used for the estimation of the matrix as well as validation. Survey data are conventionally used as an input for estimating an OD matrix. However, both sources (mobile phone data and travel survey data) have their strengths and weaknesses, possible errors, and biases. Comparison of both data sources provides an indication of the validity of the mobile-phone-data-derived OD matrix, but also the flaws of household surveys are important. Therefore, our hypothesis is that the combination of both data types improves the quality of the a priori OD matrix for transport modelling. Because of the flaws of both sources and the proposed combination of both to construct one a priori matrix, we use a different approach to assess the validity of the matrix, comparing the quality of the a priori OD matrix based on the extent in which assigning this OD matrix (without calibration) reproduces the counts in the network. Although this type of validation has been done earlier as well, we are not using the count data to calibrate the produced OD matrix nor using it for translating OD-derived trips towards absolute number of trips. We combine the knowledge and findings of earlier research to improve the estimation of an a priori OD matrix. Improving the a priori OD matrix results in a better starting or reference point for the a posteriori OD matrix estimation procedure and therefore also an improvement in predictions made. This research is done for the transport model of the Rotterdam Region, a large operational transport model in the Netherlands. In our research CDR are used to derive an a priori OD matrix, and then it was compared with travel survey data. The quality of this enriched a priori matrix was also compared to the operational (and traditional) a priori OD matrix based on survey data. Then a combination of traditional methods and mobile phone data is proposed and assessed.

In the next section, we provide background information on transport modelling and introduce the operational transport model used for the Rotterdam Region. Following that, we will describe the mobile phone data used in this research. The analysis of the mobile phone data, including a comparison with household travel surveys is provided in the following section. Then we show the enrichment procedure, using mobile phone data in combination with the operational transport model and discuss the obtained results. This paper ends with conclusions and the next steps for further research.

Transport Modelling

General

Accurate information is a premise for good decision making, as is the case in transportation. Policymakers often rely on information from transport planning models, because they provide an average picture of the current state (average working day of a current year describing mainly the number of trips between zones, used modes and routes, link flows, travel times, and congestion problems), prediction of future state and expected effects of policy measures (Mackie, 2010; Mouter, 2014; Benbow et al. 2008). Estimations from transport planning models for current as well as predicted states are also input for impact models providing, for example, information on air quality and noise (Wismans et al., 2011).

Transport planning models typically estimate demand and assign this demand (e.g., trips) to the supply of infrastructure estimating route choice and flows on every network link in four steps: production/attraction, distribution, modal split, and assignment (Ortuzar and Willumsen, 1990; Benbow et al., 2008, Bliemer et al., 2014; Van Eck et al., 2014). Figure 1 shows an overview of the traditional four-step transport model. In the production/attraction step the number of leaving and entering trips per zone is determined. The distribution step connects the productions and attractions to determine the number of trips between zones, resulting in a total a priori OD matrix. The modes used for these trips are determined in the modal split step resulting in OD matrices per mode. These first three steps are often called the demand model. In the last step, these matrices (per mode) are assigned to the network (of the corresponding mode) to determine routes and link loads using a network model. Because the demand model often uses travel times and equilibrium is assumed, iterations are needed feeding the demand model with assignment results to calibrate the a priori matrix improving the match between assignment results and counts to estimate the final a posteriori OD matrix. In practice several steps in the demand model can be combined (e.g., simultaneous distribution and modal split steps) or extended with additional steps (e.g., departure time choice).

Although transport models need input in the form of measured data, the state-of-the-art models and definitely the state-of-the-practice models still focus on an average traffic state and heavily rely on modelling assumptions using the same type of data as has been done for decades. These are mainly stated and revealed preference surveys among a sample of (potential) travelers to estimate and calibrate behavioral models and loop detector counts or incidental field investigations to calibrate the entire model (Joksimovic and Hofman, 2014).

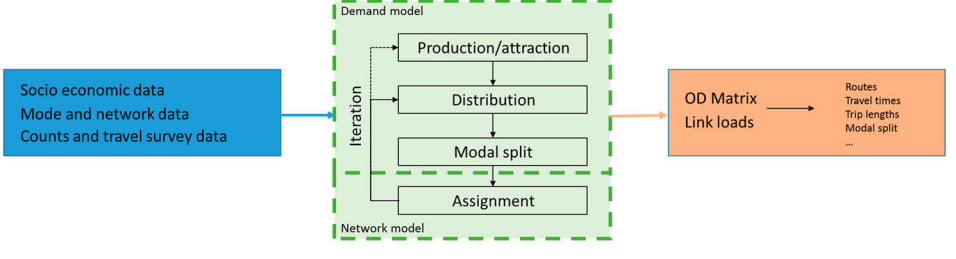

Figure 1. Traditional four-step transport model

Transport Model of Rotterdam Region

In our research, we used the state-of-the-practice transport model of the Rotterdam Region. It is a multi-modal model (car, trucks, public transport, and active modes) distinguishing nearly 6,000 TAZ (Traffic Analysis Zones) using a detailed network of the region. Although the model focuses on the Rotterdam Region, the model itself incorporates a far larger network of the Netherlands including network and zones outside the Rotterdam Region with less detail. This is needed for proper estimation of trips coming from outside the Rotterdam Region towards the Region and vice versa and also for estimation of through traffic. The a priori OD matrix of the Rotterdam region is built using demand patterns and travel characteristics that are determined based on the Dutch household travel survey database OViN (Centraal Bureau voor de Statistiek, 2017). The model is built for the modes: vehicles (cars and freight), public transport, and bikes and represents an average working day including two-hour peak periods in the morning and afternoon. Calibration of the synthetic a priori OD matrix on counts results in the a posteriori OD matrix. The model uses an equilibrium assignment based on link travel time functions and junction delays. This model is one of the largest and most detailed operational transport models in the Netherlands, and it is used for many policy decisions. Figure 2 illustrates the modelled loads per mode on a part of the network.

Although this model is state-of-the-practice, the model is mainly based on the use of traditional data sources to validate and calibrate the model. In addition, the use of a gravity model to connect productions and attractions neglects historically grown strong relations between zones that cannot be properly estimated based on the number of productions and attractions per zone and the skims (i.e., impedance between zones reflected

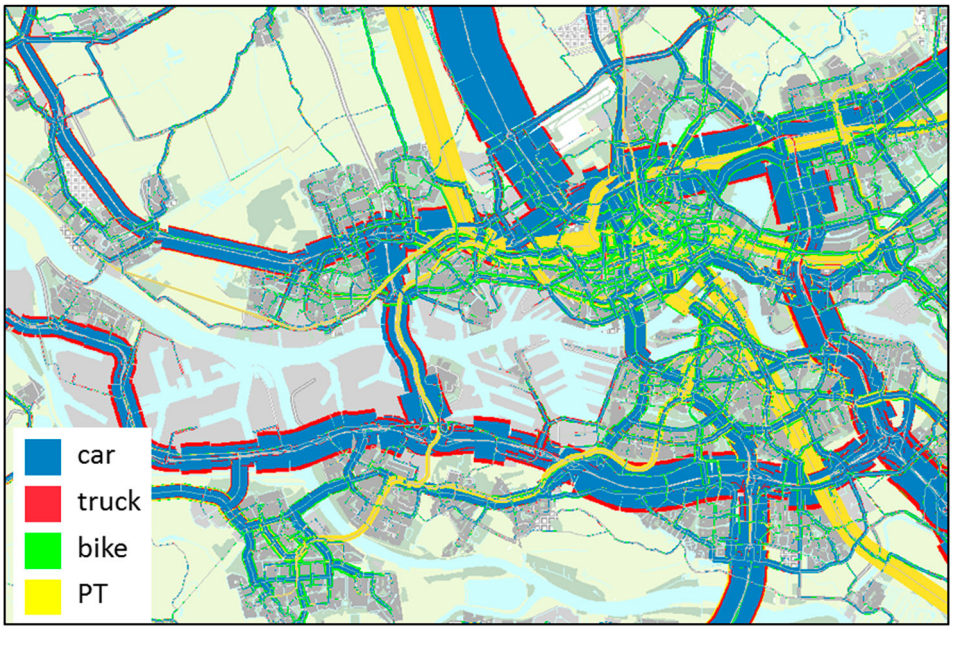

Figure 2. Loads of transport model Rotterdam Region per mode: car, truck, bike and public transport (PT). The larger the bands, the larger the amount of traffic

by the generalized costs between these zones). This is also shown by De Romph et al. (2015), in the Senegal case, but is also true for other countries like the Netherlands. In the Rotterdam Region there is for instance a much stronger relation between Zoetermeer and The Hague than would be expected from socioeconomic perspective.

Mobile Phone Dataset

Description of Dataset

The mobile phone data used are call detail records (CDR). These are records describing when a mobile phone is actively connected to the mobile phone network by sending or receiving voice or text, or when using data. The records consist of a time stamp, a cell code relating to a cell tower in the network and a one-way hashed ID created from a mobile phone number. The privacy of the raw mobile phone data is assured by a rigorous protocol. First, we do not have direct access to this raw data, but were only allowed to query this data. Second, the identifying information (phone number) is one-way hashed. The hashing is changed each first day of the month; in this way the movements of a single device can no longer be recognized after one month. Third, only groups can be extracted from the raw data (which we call the "minimal 16 rule" and is set by the mobile phone operator). We are allowed to query all raw CDR data from individual devices, but only groups of 16 devices or more are giving results in the export. In this way, it is impossible to relate information to a single device and determine individual trip information. However, by following a procedure of aggregating trips over multiple days while querying the raw data and determining the average number of trips per day after exporting these aggregated data, we are able to determine that average trip rates are below 16 per day.

Derivation of OD Information Based on Mobile Phone Dataset

The data were provided by the data processing company Mezuro. The raw data is processed into basic location information removing "noisy" data (e.g., data from mobile devices with low activity). The location is determined based on cell tower ID and the associated cell tower plans for the selected time period. The location is determined by using rule-based algorithms. The algorithms uses multiple time ordered events of single devices and thus multiple cell towers (i.e., locations) in order to increase the accuracy of determining the location. The multiple events within a timeframe are later on also used to determine if the analyzed timeframe is a "stay" or not (i.e., device was situated at the same location or was moving). The rules were created and validated in earlier research of Mezuro by using samples where the real locations were known. Because of the known issues regarding spatial accuracy of determining the location using CDR data (e.g., described in Bonnel et al., 2015), the location data are aggregated at the level of villages. As a result, the Netherlands is split into 1,259 areas for which OD-information of mobile phones is available, where each city or village is a separate zone. This zone definition is the result of earlier analysis of Mezuro of the dataset provided (CDR and cell tower plan properties). The largest cities in the Netherlands are split into city districts. For example: the city of Rotterdam is divided into eight districts. The location data are subsequently translated into OD-information based on the time-ordered stay sequence

of single devices. This translation is also done by using a rule-based approach. A mobile device is defined as a "stay" when it is at a single location for more than 30 minutes. Two consecutive stays at different locations results in an OD trip between two locations. The time period associated with this trip is set on the time period situated in the middle of the two consecutive stays (i.e., middle between end time first stay and starting time second stay). In the query, performed at the data provider behind the firewall, first data for the requested time period are selected out of the raw data. Based on the selected period and if needed the time of day the OD-information is computed and the "minimal 16 rule" is applied at the end of the query. Note that the data contain no intra-zonal information.

We selected the working days of the month of November 2014. November is a month without holiday periods in the Netherlands and therefore a representative month for the average traffic conditions during working days (24-hour level), which is also the focus of transport models. To minimize the effect of the "minimal 16 rule" the OD-data of the sum of the working days in this month were provided and divided by the number of working days to obtain the average working day (instead of averaging the data set for each working day separately). This data set contains more than 1,000,000 OD-trips. Due to the aggregation, the "minimal 16 rule" excludes on average 1 percent of the trips.

In this way, we are equipped with OD-information of mobile phone devices (total number of trips) for the Netherlands at the level of villages for the month of November 2014. First, we started by analyzing the dataset to determine characteristics of the data and in what way we can incorporate this OD-information in transport models. Based on the results of this analysis it was concluded that the OD-information is very promising.

With a transport model distinguishing nearly 6,000 zones and OD information based on mobile phone data distinguishing 1,259 zones, there is a mismatch in zone detail. This means that we can use the data in the enriching procedure only for external trips as is seen in Huntsinger and Ward (2015). In Figure 3, the OD-data of mobile phone data for a single day is presented showing a plausible spatial distribution. The figure shows that most trips are made in the Randstad area (Amsterdam, The Hague, Rotterdam and Utrecht).

Data Analysis

Visual Inspection

Figure 3 showed that the observed trips based on mobile phone data are concentrated between and near cities. Visual inspection of the trips shows that the spatial distribution is generally plausible. Figure 4 shows for example the number of trips originating in the city of Rotterdam. The figure shows, as expected, trips in all directions nearby but also with the larger cities further away like Amsterdam, Utrecht, and Dordrecht (which is indeed strongly oriented on the city of Rotterdam).

Even though the transport model focuses on the Rotterdam region, the productions and attractions of areas outside this region are also needed, because they influence the number of trips as well as the routes used towards Rotterdam and vice versa. This means that other typical relations within the Netherlands are of importance in building a valid model for the Rotterdam Region. The trips derived from the mobile phone data also clearly show other

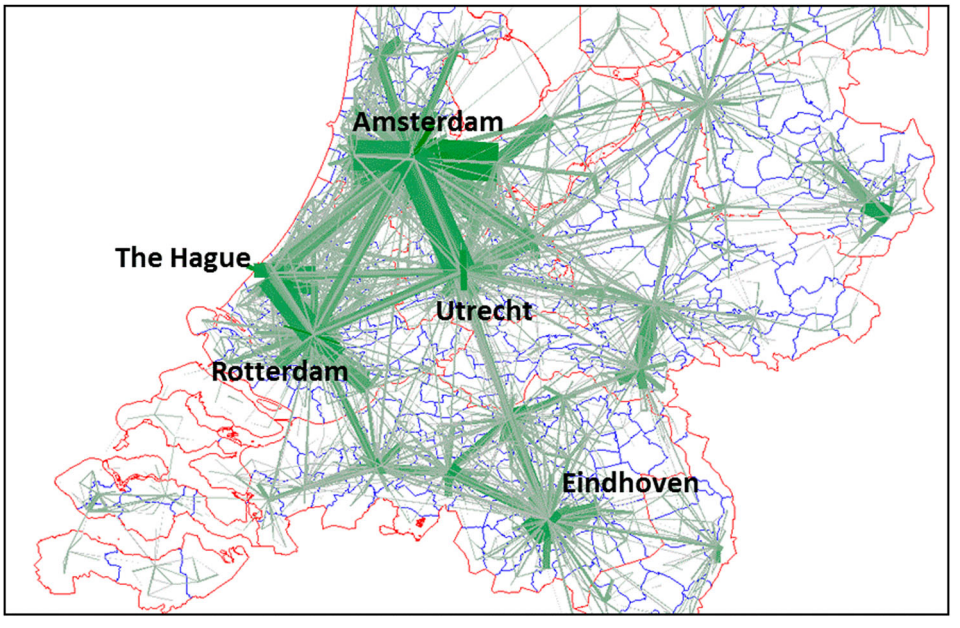

Figure 3. Spatial distribution of trips from mobile phones

typical relations being present in the mobile phone data set. Figure 5 shows, for example, the distribution of trips with origin in the city of Almere. Once originated as an Amsterdam suburb, nowadays Almere is the seventh largest city of the Netherlands with about

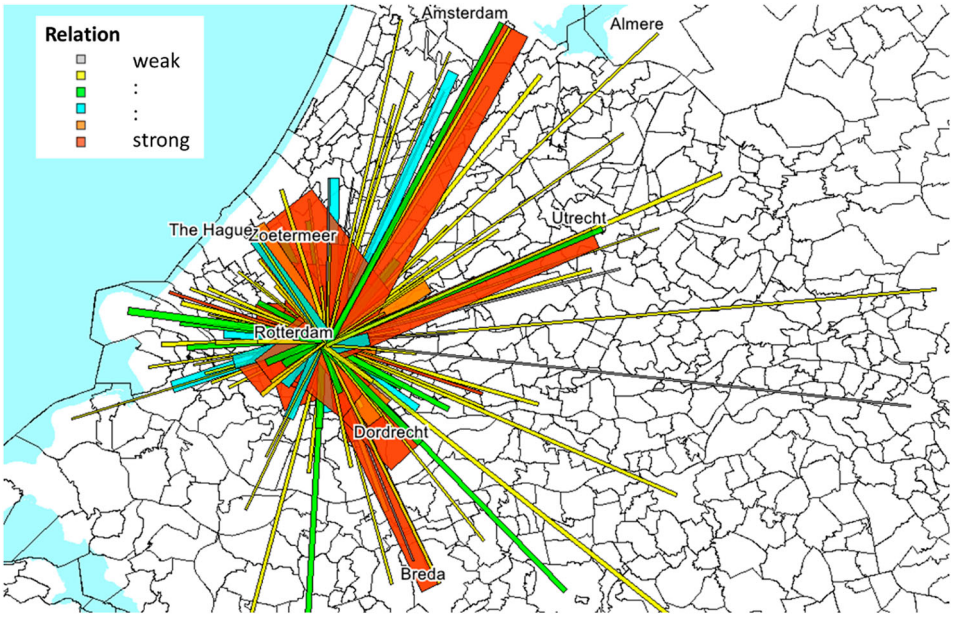

Figure 4. Distribution of trips originating in the city of Rotterdam determined from mobile phone data (coloring and bandwidth indicate relative number of trips)

200,000 inhabitants. Many of them are still strongly related to Amsterdam in terms of commuting and family visits. This phenomena is clearly visible in Figure 5.

The same phenomena is visible in Figure 6. Zoetermeer originated as a suburb of The Hague. From a historic perspective a lot of commuters living in Zoetermeer work in The Hague, which is clearly present in the mobile phone data. In a traditional transport model, it is hard to model these kinds of historic relations, which clearly show the added value of incorporating mobile phone data in transport modelling. For example, the city of Leiden is about 15 kilometers (north) of Zoetermeer, as is The Hague, and a traditional transport model would allocate as many trips from Zoetermeer to The Hague as from Zoetermeer to Leiden.

However, the data also shows a small number of long-distance trips generated in the Harbor area of Rotterdam, which would be expected given the high number of trucks using this route. Although we cannot investigate the raw CDR data, it is expected that this can be explained because of the 30-minute criterion and privacy filter of at least 15 trips. The number of truck-trips between a specific origin and destination are likely to be less than 15 in a lot of cases. Taking into account that many trucks will also travel to other countries and the loading and unloading of trucks takes less than 30 minutes on some occasions, it is also possible that some destinations are not identified as a destination. As a result, it can be expected that truck trips are underrepresented in the mobile phone data set.

Comparison with Household Surveys

The mobile phone data set contains a large set of approximately one-third of the total Dutch population, but is not complete and probably biased. In this research we only

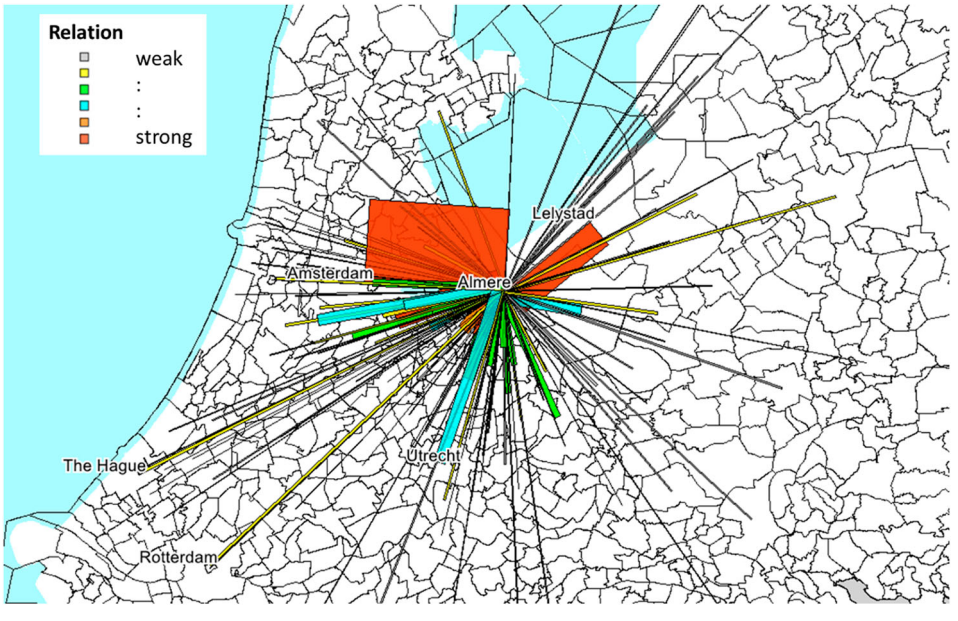

Figure 5. Distribution of trips originating in the city of Almere determined from mobile phone data (coloring and bandwidth indicate relative number of trips)

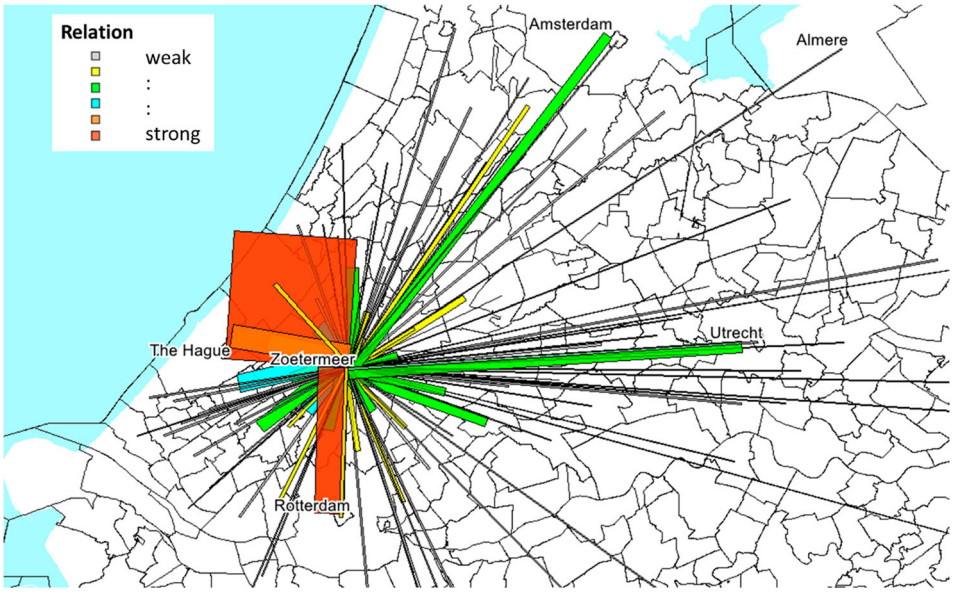

Figure 6. Distribution of trips originating in the city of Zoetermeer determined from mobile phone data (coloring and bandwidth indicate relative number of trips)

used the distribution of trips, which means that it was not necessary to increase the measured trips to produce absolute values. However, also this distribution is probably biased. To investigate this, we performed a comparison analysis between characteristics of the mobile phone data and the Dutch household travel survey OViN (Onderzoek Verplaatsingen in Nederland) (Centraal Bureau voor de Statistiek, 2017). Every year, this survey selects a representative sample of Dutch households who are asked to fill in a travel diary for two or three days. In this diary, they report which trips were made in their household, by what mode and what kind of purpose, resulting in approximately 35,000 filled in questionnaires. The OViN is a database of big importance in current transport modelling in the Netherlands, using demand patterns and travel characteristics. However, as shown by Stopher and Greaves (2007), these types of travel surveys also contain a possible bias and errors in the number of reported trips. It is expected that the bias in trip length distribution of mobile phone data is larger than for OViN data. There are several reasons for this. OViN data are based on a representative sample of Dutch households but this is not the case with mobile phone data. There is a bias related to the use of mobile phone data, and location errors occur because of an expected bias for short trips. Also, it is important to know that each source investigates different kinds of trips. One main difference is that trips made by trucks are not part of OViN. In addition, the sample size of OViN (35,000 annually) is much smaller than the mobile phone data set (>1 million monthly), so that the variance in the mobile phone data is expected to be smaller.

In the comparison of trip length (See Figure 7) it can be seen that as expected mainly the number of short distance trips (shorter than 8 km) are underrepresented in the mobile phone data set. Above 8 km the mobile phone data set is overrepresented, which slowly decreases and is almost dissolved at 60 km. In Figure 7 the trip length

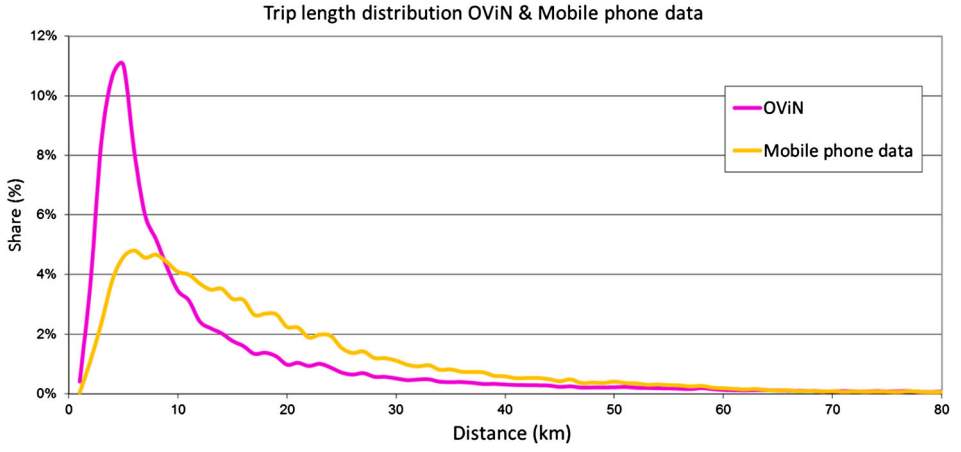

Figure 7. Comparison trip length (crow flies) distribution of mobile phone data and OViN

distribution frequencies sum to 100 percent, which means that underrepresentation in one part of the graph automatically leads to over-representation in another part. Analysis of the distribution focusing on the range between 8 and 60 km shows that the trip length distributions from about 40 km correspond reasonably well. The range of 8 to 15 km is still significantly underestimated, while the range between 15 and 40 km is slightly overestimated.

The underrepresentation of short-distance trips can be the result of less mobile phone use on short distances or that people did not take their mobile phones with them. Another explanation is also that the possibility to measure a trip based on the raw mobile phone data depends on the spatial resolution of cell towers and the activity duration at the destination. The number of trips is decreasing by increasing distance and therefore there is a probability that long-distance trips based on CDR data will be filtered out because of the privacy definition ("minimal 16 rule"). This means that the probability that more than 15 trips are made between a specific origin and destination zone is smaller for larger distances.

Enrichment A Priori Demand Estimates

Taking into account that mobile phone data do not contain intra-zonal trips and have less detail than the transport planning model of the Rotterdam regions as well as the fact that we only have a sample of the total number of trips, results in the following procedure of the enrichment of the a priori demand estimates using mobile phone OD information:

- determine the average working-day OD matrix from the mobile phone data
- convert the model zoning system (about 4,000 zones) to the zoning system of the mobile phone data.
- aggregate of the a priori OD matrices per mode (car, PT, and bicycle) and per purpose to determine a working day personal transport OD matrix
- scale the synthetic a priori OD-relations of the model (row & column totals) to the measured OD-relations from the mobile phone data

- expand the enriched OD matrix to the zoning system of the Rotterdam model as well as the aggregated modalities, using the splitting rates according to the a priori model resulting in an enriched a priori OD matrix per mode.

The diagonal of the a priori OD matrix is not taken into account in the enrichment procedure because there is no intra-zonal information available in the mobile phone data. Next to that, the a priori truck matrix has not been part of the enrichment procedure using mobile phone data because of the identified underrepresentation of these trips. In the second step—"converting the model zoning system"—it has been taken into account that in the study area of Rotterdam, the model zones are geographically smaller than the mobile phone zones. Therefore, in this part the model zones are aggregated. While in the so-called influence and outer area of Rotterdam, the zoning system of the mobile phone data is more detailed than the model zoning system. Therefore, in this part of the Netherlands the data of the mobile phone zones are aggregated. In the fourth step—"scaling of the synthetic a priori OD-relations"—the distribution of the synthetically determined OD-information (traditional a priori model based on survey data) is replaced by the distribution of the measured OD-information from mobile phone data, while maintaining the absolute demand estimated in the a priori matrix. To determine the new OD information, the Furness procedure is used, which is a standard procedure often used to ensure that the summation of individual cells results in the fixed row and column totals while retaining the distribution inserted.

We performed this enrichment procedure using two different approaches:

Approach 1: Scaling based on the distribution of the mobile phone data for all distances

Approach 2: Scaling based on a corrected distribution taking into account the found bias in trip-length distribution. The correction is based on the comparison between OViN data and mobile phone data (See Figure 7). As a result trips shorter than 8 km and longer than 60 km were not enriched using mobile phone data (i.e., the a priori estimated trips in the transport model are maintained). Furthermore the trip distribution for trips between 8 and 60 km are corrected based on the differences found in the total trip length distribution between OViN data and mobile phone data. As a result the number of trips between 8 and 13 km are increased, between 14 and 40 km decreased and larger than 40 km unchanged. Note, that on individual relations this does not result in the same distribution.

Results

Assessment Framework

The enrichment procedure is tested on the operational transport planning model of the Rotterdam region. Because the ground truth of demand is not known and a combination of data sources is used for the enrichment procedure, the assessment is based on comparing the assignment results using the original a priori OD matrix estimated in the model with the assignment results of the enriched a priori OD matrix using mobile phone data for both approaches. For all cases, a comparison is made of the fit between the assignment results with count data. Note that none of the matrices are calibrated against count data. As a result, the comparison shows to what extent these matrices are capable of

reproducing counts. Furthermore, we also assessed the results using the expert knowledge of local traffic experts of the municipality of the mobility patterns in the Rotterdam region.

First analysis focused on the model fit. The model fit is defined by determining the so-called T-value at each count location, comparing the modelled value (flow) with measured value (count). In total there are 1,180 locations for which counts are available. The T-value (not to be confused with the T-statistic) at location c is defined as follows:

$$T_c = \ln\left(\frac{(X_c - X_{m,c})^2}{X_c}\right) \tag{1}$$

where X_c is the count value and $X_{m,c}$ the modelled traffic flow at location c. The T-value not only takes the absolute difference into account, but also the absolute count value. Using this measure, locations with a high count value (like motorways) can be rated in the same way as locations with a low count value (like a local road). If $T_c <= 3.5$ this is rated as no relevant deviation; if $3.5 < T_c <= 4.5$ this is rated as a border case and a value $T_c > 4.5$ as a relevant deviation.

Second, we analyzed the correlation coefficient between the counts measured versus the modelled flows as well as the root mean square error (*RMSE*). Because this is a comparison on an aggregated level, we also analyzed the RMSE of the 10 percent number of locations with the highest deviations ($RMSE_{10}$) of the a priori model. This last value indicates to what extent the enriched a priori model is capable to improve the general structure of the a priori matrix (i.e., reducing deviation outliers representing difficult-to-reproduce counts). Better performance on model fit (regarding the T-values), correlation coefficient as well as $RMSE_{10}$ indicates an improvement in OD matrix quality. Furthermore this also indicates that the influence of the calibration towards the a posteriori matrices decreases, improving the predictive power of the transport model.

Assessment Based on Assignment Results

Analysis of Model Fit. The result of the assignment of the model is assessed for 1,180 count locations. Table 1 shows the performance of the assignments compared with the traffic counts of the original a priori matrix and the two enrichment approaches. The earlier mentioned finding that the bias in Approach 1 is too important to neglect, is clearly shown when compared with the a priori model result. The number of count locations with a T-value below 3.5 (e.g., no relevant deviation) has decreased by 45 (392 vs. 437). Approach 2, however, shows an improvement of approximately 5.4 percent in T values below 3.5 when compared with Approach 1. This improvement means that 1.6 percent more of the counts are rated as no relevant deviation (in comparison with the a priori model fit) (See Table 1).

Table 1. Analysis of model fit: comparison of link loads with count values

Range T-value	A-priori model		Enriched: Approach 1		Enriched: Approach 2	
$T_c < 3.5$	437	37.0%	392	33.2%	456	38.6%
$3.5 < T_c <= 4.5$	235	19.9%	171	14.5%	228	19.3%
$T_c > 4.5$	508	43.1%	617	52.3%	496	42.0%
Number of counts	1,180		1,180		1,180	

Table 2. Analysis of correlation coefficient

	Approach		
Indicator	A priori model	Enriched: Approach 1	Enriched: Approach 2
Correlation coefficient	0.955	0.954	0.959

Analysis of Correlation Coefficient. Table 2 presents the results of the three approaches for the correlation coefficient. The correlation coefficient shows small differences, which could be expected based on the analysis of the model fit and the relatively large number of counts with small values. Approach 1 shows a small deterioration and Approach 2 shows a small improvement compared to the original a priori model.

Analysis of RMSE and RMSE$_{10}$. The scatter plots in Figure 8 show the deviation for all locations (measured flow versus modelled flow) for the three approaches. The plots show that Approach 1 performs worse compared to the original a priori, and Approach 2 shows an improvement, especially for the counts with higher values containing the major roads in the network. Therefore, we also analyzed the RMSE and RMSE10 performance (See Table 3), clearly indicating that Approach 2 is performing better than a priori (–13 percent for RMSE and –18 percent for RMSE$_{10}$), but also that Approach 1 performs worse (+90 percent for RMSE and + 76 percent for RMSE$_{10}$) than a priori.

These results show that calibration effects will be reduced by calibrating the enriched model of Approach 2 instead of the synthetic a priori model. In building a model for a forecast year also, the determined calibration effects are used and this implies that the quality of the forecast also will improve. Furthermore, it also shows what happens when only the OD-information based on mobile phone data is used without addressing the bias results in a deterioration of the transport model.

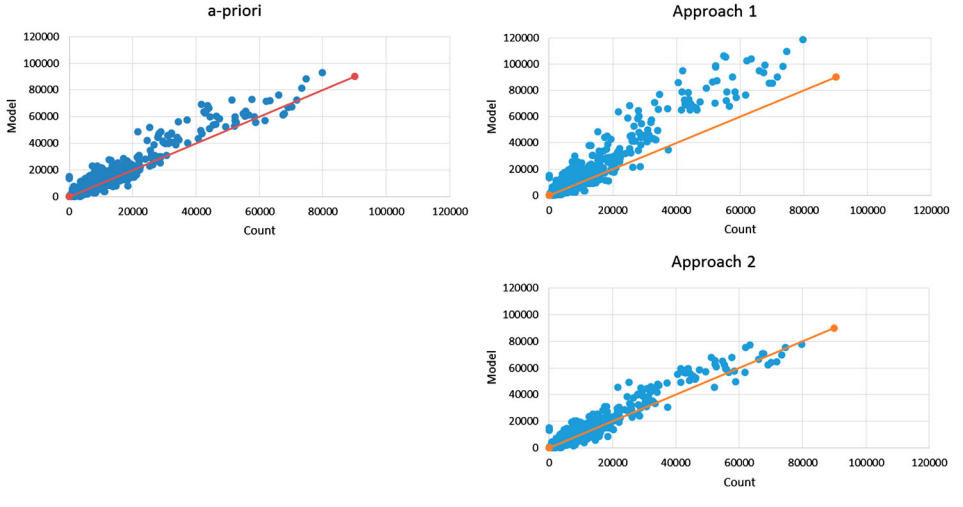

Figure 8. Scatter plots model values versus counts for three approaches

Table 3. Analysis of RMSE and $RMSE_{10}$

	Approach		
Indicator	A priori model	Enriched: Approach 1	Enriched: Approach 2
RMSE	4,424	8,409	3,838
$RMSE_{10}$	10,426	18,358	8,512

Assessment Based on Expert Judgment

The results were also discussed with experts of the Rotterdam municipality. Approach 1, in which the bias in trip length distribution was neglected, resulted—as expected—in an overestimation of long trips in the model and large changes in the structure of the original a priori matrix. This is due to the underrepresentation of short trips when applying the distribution based on mobile phone data. Overestimation of long trips results after the assignment step to large overestimations of link loads especially on higher order roads. The enriched a priori OD matrix based on Approach 1 results therefore in a deterioration of the original matrix. Therefore, the results of Approach 1 were not discussed with experts.

Approach 2, taking into account the bias found, provides promising results, which have been analyzed further. Visual analysis shows that most of the link loads are near the measured counts. A comparison of both assignments (i.e., original a priori versus enriched a priori OD matrix) is indicated in Figure 9, showing the region between The Hague (in the north-west) and Rotterdam (in the south). In this figure the bandwidths show the traffic flows, where red parts mean a higher traffic flow in the assignment of the enriched OD matrix and green parts mean a higher traffic flow of the a priori model.

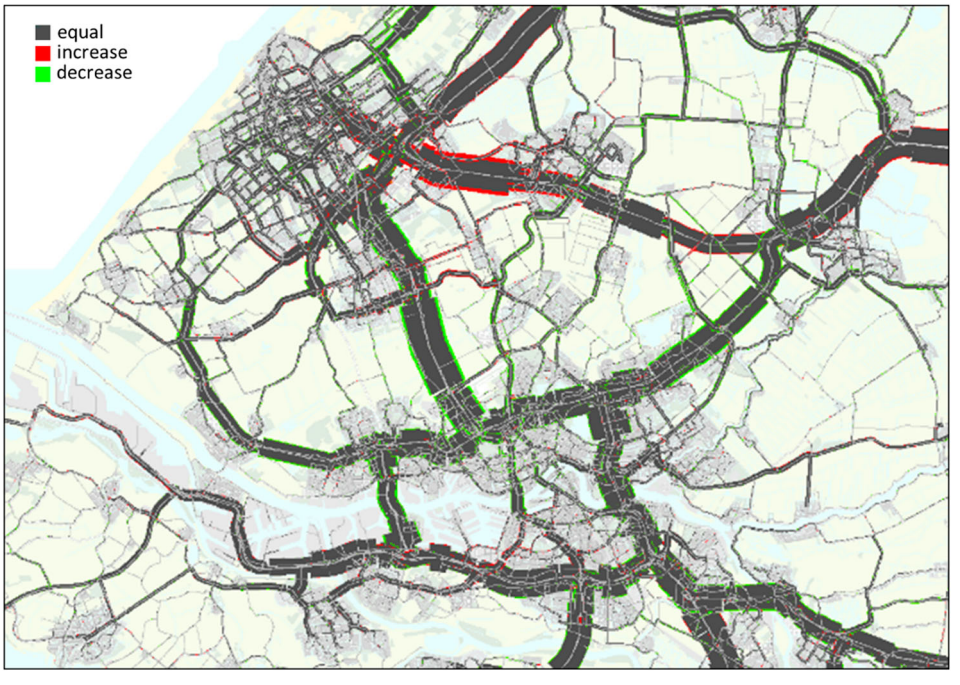

Figure 9. Comparison of car assignments of: enriched a priori (Approach 2) versus original a priori

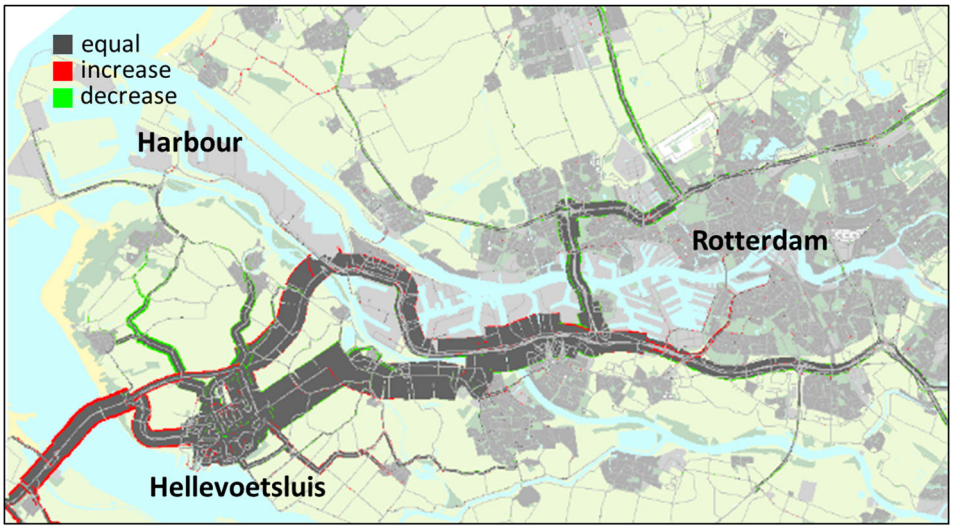

Figure 10. Comparison of selected area assignment of car in the city of Hellevoetsluis

Figure 9 also shows the earlier presented historically grown strong relation between Zoetermeer and The Hague, which is difficult to model using the traditional approach (See also Figure 6). Further analysis has been performed on the city of Hellevoetsluis, south-west of the city of Rotterdam. It is known that the distribution of trips related to this city always has been hard to model in the Rotterdam transport model, because of the isolated geographical position of the city and the nearby Harbor. The gravity model expects a much stronger relation between Hellevoetsluis and the Harbor area than is present in reality. Figure 9 shows that the traffic flows near Hellevoetsluis hardly have changed, but in Figure 10 it becomes clear that the distribution has changed. The figure shows all trips with an origin and/or destination in Hellevoetsluis. These changes are consistent with the experience of regional traffic engineer experts.

Conclusions and Further Research

Mobile phone data have a great potential to improve current transport modelling. Further development of methods and combining different data sources will lead to more data-driven modelling approaches used in practice. Although there have been previous efforts on deriving OD matrices from mobile phone data, validation, when using real mobile phone data, was often missing or done by comparison with household survey data. However, because mobile phone data are associated with a bias (that is also the case for traditional household survey data), we proposed combining the strengths of traditional a priori OD matrix estimations based on survey data and the OD information derived from mobile phone data. We used traditional a priori matrix to provide absolute trip levels, combining trip length distributions based on travel survey data and mobile phone data to improve the structure of this traditional a priori matrix. Furthermore, we proposed evaluating the assignment results of the estimated a priori matrix, which was not done in earlier research. This means we did not use counts to calibrate our a priori matrix, making sure that the calibration process could not influence the results. The

results showed that it is possible to improve the quality of the a priori matrix when addressing the bias when combining both data sources. The OD information derived from the mobile phone data showed an underrepresentation of short trips, which is partly due to the level of detail and accuracy of the available location data for this research. This confirms earlier research findings. When this bias is not addressed, the use of OD information from mobile phone data leads to a deteriorating performance. The method proposed combining the data correcting for the bias regarding short trips showed an improvement in assignment results comparing count data with modelled flows. This means we estimated a better performing a priori OD matrix, which means that the calibration effects will be smaller. Smaller calibration effects indicate an improvement in the prediction power of this kind of transport model. In comparison with earlier literature, we showed that the combination of traditional methods with mobile phone data to estimate an OD matrix results in an improvement and not only to similar results when comparing traditional methods with a mobile-phone-data-derived OD matrix. Further research on combining the strengths of several data sources could result in major improvements in transport modelling.

Further improvements are possible when it is possible to improve the spatial resolution of the OD-information derived from mobile phone data as well as improvement of detection of small numbered trips as is indicated by earlier research as well. Further research is needed on the bias associated with the data as well as the performance for smaller time periods (i.e., peak-hour models versus 24-hour models) and ways to correct for this bias. Furthermore, there is a need for appropriate validation data. Although the use of assignment results compared to counts provides an objective assessment, there is an influence of the route choice model used on the outcomes. Other research already showed that it is possible to derive trip purposes, which could further improve the enrichment procedure. Next to that, further research is needed in proper estimation of mode choice, providing the possibility of improving the OD matrix per mode. Combining mobile phone data with other sources for this purpose can be a solution. Finally, the data also show great potential to include day-to-day dynamics in transport modelling, improving the assessment of measures in practice. Research will be needed to incorporate this.

Disclosure Statement

No potential conflict of interest was reported by the authors.

Notes on Contributors

Luc Wismans is an associate professor at the Centre of Transport Studies, University of Twente and program manager of Smart Mobility at the Goudappel Group.

Klaas Friso is a senior consultant at DAT.Mobility. He works on integrating Big Data sources in transport models.

Jeroen Rijsdijk is a data scientist and transport modeling expert at the municipality of Rotterdam. He works on the development of transport models and the deployment of data to support policy decisions.

Stefan de Graaf is a senior transport modeler at Goudappel Coffeng. He has developed the transport model of the Rotterdam region commissioned by the municipality.

Jasper Keij is a data scientist at Mezuro. He works on the algorithms deriving information from Call Detail Records.

Bibliography

L. Alexander, S. Jiang, M. Murga, and M.C. Gonzalez, "Origin-Destination Trips by Purpose and Time of Day Inferred from Mobile Phone Data," *Transportation Research Part C* 58 (2015) 240–250.

H. Bar-Gera, "Evaluation of a Cellular Phone-Based System for Measurements of Traffic Speeds and Travel Times: A Case Study from Israel," *Transportation Research Part C* 15 (2007) 380–391.

N. Benbow, P. Kidd, A. Woolley, A. Skinner, and I. Palmer, "The Continued Innovation of Aggregate Transport Demand Models," paper presented at European Transport Conference (Leeuwenhorst Conference Centre, Noordwijkerhout. The Netherlands, 6–8 October 2008).

M.C.J. Bliemer, M.P.H. Raadsen, L.J.N. Brederode, M.G.H. Bell, and L.J.J. Wismans, "A Unified Framework for Traffic Assignment: Deriving Static and Quasi Dynamic Models Consistent With General First Order Dynamic Traffic Assignment Models," paper presented at DTA 2014 symposium (Salerno, Italy, 17–19 June 2014).

P. Bonnel, E. Hombourger, A.M. Olteneanu-Raimond, and Z. Smoreda, "Passive Mobile Phone Dataset to Construct Origin-Destination Matrix: Potential and Limitations," *Transport Research Procedia* 11 (2015) 381–398.

N. Caceras, J.P. Wideberg, and F.G. Benitez, "Deriving Origin-Destination Data from a Mobile Phone Network," *IET Intelligent Transport Systems* 1: 1 (2007) 15–26.

F. Calabrese, G. Di Lorenzo, L. Liu, and C. Ratti, "Estimating Origin-Destination Flows using Mobile Phone Location Data," *IEEE Pervasive Computing* 10: 4 (2011) 36–44.

Centraal Bureau voor de Statistiek (CBS); Rijkswaterstaat (RWS), Onderzoek Verplaatsingen in Nederland 2016 - OViN 2016. DANS. (2017) <https://doi.org/10.17026/dans-293-wvf7> Accessed February 1, 2017.

S. Çolak, L.P. Alexander, B.G. Alvim, S.R. Mehndiretta, and M.C.G. Gonzalez, "Analyzing Cell Phone Location Data for Urban Travel: Current Methods, Limitations and Opportunities," paper presented at TRB conference 2015 (Washington, USA, January 11–15, 2015).

E. de Romph, G.H.A. Correia, and Y. Wang, "National and Regional Road Network Optimization for Senegal Using Mobile Phone Data," Data for Development Challenge (D4D) paper presented at Netmob Conference 2015 (Cambridge, USA, 9–10 April, 2015).

D. Gundlegard, C. Rydergren, N. Breyer, and B. Rajna, "Travel Demand Estimation and Network Assignment Based on Cellular Network Data," *Computer Communications* 95 (2016) 29–42.

L.F. Huntsinger and K. Ward, "Using Mobile Phone Location Data to Develop External Trip Models," paper presented at TRB conference 2015 (Washington, USA, January 11–15, 2015).

S. Iqbal, C.F. Choudhury, P. Wang, and M.C. Gonzalez, "Development of Origin-Destination Matrices Using Mobile Phone Call Data," *Transportation Research part C* 40 (2014) 63–74.

D.K. Joksimovic and F. Hofman, "The Recent Developments of the Dutch Regional and National Models," paper presented at European Transport Conference (Frankfurt, Germany, 29 September–1 October, 2014).

J. Ma, H. Li, F. Yuan, and T. Bauer, "Deriving Operational Origin-Destination Matrices From Large Scale Mobile Phone Data," *International Journal of Transportation Science and Technology* 2: 3 (2013) 183–204.

P. Mackie, "Cost Benefit Analysis in Transport: A UK Perspective," Discussion paper, no 2010-16, Prepared for the OECD/ITF Round Table, on Improving the Practice of Cost Benefit Analysis in Transport (Mexico, 21–22 October 2010).

N. Mouter, *MKBA Internationaal. Lessen uit een Vergelijking van de Nederlandse MKBA-Praktijk met Vier Andere MKBA-Praktijken*. [In Dutch, translated: Cost Benefit Analysis Internationally, Lessons Learned Comparing Dutch Practice with Four Other Practices]. Commissioned by Planbureau voor de leefomgeving (PBL) (2014) Delft.

M. Nanni, R. Trasarti, B. Furletti, L. Gabrielli, P. Van Der Mede, J. De Bruijn, E. de Romph, and G. Bruil, "MP4-A Project: Mobility Planning For Africa. D4D 2013 – Data for Development," paper presented at Special session of the Third International Conference on the Analysis of Mobile Phone Datasets (Cambridge, USA, 2–3 May 2013).

J.D.D. Ortúzar, J. Armoogum, J-L. Madre, and F. Potier, "Continuous Mobility Surveys: The State of Practice," *Transport Reviews* 31: 3 (2011) 293–312.

J.D.D. Ortúzar and L.G. Willumsen, *Modelling Transport* (Chichester: Wiley, 1990).

P.R. Stopher and S.P. Greaves, "Household Travel Surveys: Where Are We Going?" *Transportation Research Part A* 41 (2007) 367–381.

R. Tolouei, S. Psarras, and R. Prince, "Origin-Destination Matrix Development: Conventional Methods Versus Mobile Phone Data," *Transport Research Procedia* 26 (2017) 39–52.

G. Van Eck, T. Brands, L.J.J. Wismans, A.J. Pel, and R. van Nes, "Model Complexities and Requirements for Multimodal Transport Network Design: Assessment of Classical, State-of-the-Practice, and State-of-the-Research Models," *Transportation Research Record: Journal of the Transportation Research Board* 2429 (2014) 178–187.

M. Von Morner, "Application of Call Detail Records: Chances and Obstacles," *Transport Research Procedia* 25 (2017) 2233–2241.

J. White and I. Wells, "Extracting Origin Destination Information from Mobile Phone Data," paper presented at Road Transport Information and Control, Conference publication number 486 (London, UK, 19–21 March 2002).

L.J.J. Wismans, E.C. van Berkum, and M.C.J. Bliemer, "Modelling Externalities Using Dynamic Traffic Assignment Models: A Review," *Transport Reviews* 31: 4 (2011) 521–545, doi:10.1080/01441647.2010.544856

Z. Zhao, S.-L. Shaw, Y. Xu, F. Lu, J. Chen, and L. Yin, "Understanding the Bias of Call Detail Records in Human Mobility Research," *International Journal of Geographical Information Science* 30: 9 (2016) 1738–1762.