INVESTIGATING POTENTIAL TRANSIT RIDERSHIP BY FUSING SMARTCARD AND GSM DATA

Karin de Regt
Delft University of Technology
Faculty of Civil Engineering and Geosciences
Transport, Infrastructure and Logistics
P.O. Box 5048
2600 GA Delft, The Netherlands
Telephone: +31 6 45831801
E-mail: kldere gt@gmail.com

Oded Cats
Delft University of Technology
Faculty of Civil Engineering and Geosciences
Transport & Planning
P.O. Box 5048
2600 GA Delft, The Netherlands
Telephone: + 31 15 2781384
E-mail: O.Cats@tudelft.nl (corresponding author)

Niels van Oort
Delft University of Technology / Goudappel Coffeng
Faculty of Civil Engineering and Geosciences
Transport & Planning
P.O. Box 5048
2600 GA Delft, The Netherlands
Telephone: +31 6 15908644
E-mail: N.vanOort@tudelft.nl

Hans van Lint
Delft University of Technology
Faculty of Civil Engineering and Geosciences
Transport & Planning
P.O. Box 5048
2600 GA Delft, The Netherlands
Telephone: +31 15 27 85061
E-mail: J.W.C.vanLint@tudelft.nl

Word count:
Text (4953) + Figures/Tables (7*250) = 6703

ABSTRACT

The public transport industry faces challenges to cater for the variety of mobility patterns and corresponding needs and preferences of passengers. Travel habit surveys provide information on the overall travel demand as well as its spatial variation. However, it often does not include information on temporal variations. By means of data fusion of smartcard and Global System for Mobile Communications (GSM) data, spatial and temporal patterns of public transport usage versus the overall travel demand are examined. The analysis is performed by contrasting different spatial and temporal levels of smartcard and GSM data. The methodology is applied to a case study in Rotterdam, the Netherlands, to analyze whether the current service span is adequate. The results suggest that there is potential demand for extending public transport service span on both ends. In the early mornings, right before transit operations are resumed, an hour-on-hour increase in visitor occupancy of 33-88% is observed in several zones, thereby showing potential demand for additional public transport services. The proposed data fusion method showed to be valuable in supporting tactical transit planning and decision making and can easily be applied to other origin-destination transport data.

1. INTRODUCTION

Both passengers and the government demand an efficient public transport system with high quality and low costs. This system has to be user-oriented, and live up to the needs and preferences of the passengers *(1)*. Passengers, however, do not all have the same mobility patterns and corresponding needs and preferences. Travel demand varies not only in space, but also in time, leading to a diverse and dynamic environment *(2,3)*. To design public transport services in this dynamic environment, smartcard data are often used to analyze mobility patterns *(4).* These data, however, only provide information on the public transport travel demand, neglecting the overall travel demand although it should be taken into account by public transport operators *(5)*. Travel habit surveys are traditionally used to collect data for estimating and analyzing the demand for transport *(6,7).* Travel habit surveys are used to analyze passenger demand and preferences per modality, journey purpose and travel attributes *(6,8).* Collecting travel household survey data is a time-intensive and costly undertaking, primarily due to the labor intensive process of acquiring and processing the surveys. As a result, the surveys are performed with long intervals measured in years, aiming to represent an average (working) day for travellers *(9)*. It is therefore not possible to distinguish temporal dynamics, since only an average day is represented. This calls for the development of methods designed to acquire information on both spatial and temporal dynamic mobility patterns of public transport passengers in relation to the overall travel demand.

In addition to smart card data and travel habit surveys, several other data sources are used to gain information on mobility patterns and improve the public transport design. Examples of these data sources include automatic vehicle location systems (AVL), Wi-Fi and Bluetooth signals, social media and Global System for Mobile Communications (GSM) (*10*). The most important challenge is to process the data so that it becomes useful for improving public transport design. While AVL allows monitoring fleet performance, it does not provide information on service effectiveness. Wi-Fi, Bluetooth and social media data are only recently being used to analyze transport. These data sources offer information from a small sample of the population in high resolution and in the case of social media require complicated semantic analysis *(10)*. Therefore, these data sources do not provide information on the overall travel demand, but rather complementary information. GSM data are also increasingly used for analyzing transport demand. The extent to which GSM data are available and at which spatial and temporal level they are provided varies considerably from country to country. GSM data, are extracted from call-detail records that are provided by the network provider *(11).* Three main applications of GSM data in transport research are Origin-Destination estimation, detection of events based on crowdedness and travel mode identification *(12,13,14).* The latter is not yet applicable for GSM data in the Netherlands. Therefore relying solely on GSM data is not sufficient for the purpose of this study.

The combination of data sources, data fusion, offers a promising avenue for gaining information on public transport mobility patterns versus the overall temporal and spatial travel demand. Several data fusion studies considered either smart card data or GSM data with travel habit surveys, to successfully estimate trip purposes *(7,15).* A pilot data fusion study was performed in Emmen, the Netherlands *(16)* where smartcard and GSM data were fused to find areas with potential to provide additional public transport. A study in Singapore also explored the combination of smartcard and GSM data to identify weak public transport connections *(17).* Both studies support the hypothesis that data fusion of smartcard and GSM data offers synergies resulting with new information *(16).* Smartcard data provide information on the public transport passengers travelling with a specific operator. GSM data provide information on overall travel demand (based on a very large sample and a growth factor algorithm to scale the sample to the total population), and its temporal and spatial variation. All transport modalities are included, but no distinction can be made between the different modalities. Both data sources contain information on spatial and temporal variations.

The objective of this study is to analyze the potential of fusing smartcard and GSM data for gaining information on public transport mobility patterns versus the overall travel demand when accounting for their spatial and temporal variations. The analysis approach can be used for a variety of purposes. Many public transport operators offer special night networks and need to determine the transition times *(18)*. Also, the demand for transport throughout the night has to be examined, such that the network design and service span live up to the demand for transport during the night. We apply the data fusion analysis approach to a case study in Rotterdam, the Netherlands, in which the public transport usage versus the overall travel demand is analyzed for the late evenings and early mornings for different types of days. The aim is to identify whether the current supply of urban public transport services is adequate for the demand for transport during these hours, and in what way this varies for different types of days. The results of this study can support decision makers in evaluating the current service design and schedule and identify potential improvements.

The outline of the paper is as follows: the next section explains the proposed data fusion methodology. Section 3 applies the methodology to the case study of the night services in Rotterdam. Finally, Section 4 discusses the findings and recommendations for future applications.

2. METHODOLOGY FUSION GSM AND SMART CARD DATA

In the section, an overview of the methodology is given, starting with an overview of the structure, after which the different steps are discussed in more detail. The proposed analysis approach is illustrated in Figure 1. The main input is anonymized smartcard and GSM data along with the relevant spatial and temporal scope. Depending on the application of interest, a base case scenario is defined (e.g. representing conditions on an average day or referring to a moving reference level such as the previous hour). Input data pre-processing consists of two aspects: identification of characteristics, limitations and assumptions of each dataset, and processing the data into a workable format. Afterwards, the data fusion can be established. Hereby first different spatial and temporal analysis levels are identified by means of aggregation or differentiation in space and time. Per dataset and per analysis level the discrepancies of scenarios with respect to the base scenario are measured using quantitative metrics. The actual data fusion is established by relating the discrepancies of the smartcard data with the discrepancies of the GSM data per scenario and analysis level as explained in the following. The approach proposed in this study can be used to explore various datasets that contains information on origins and/or destinations in transport networks.

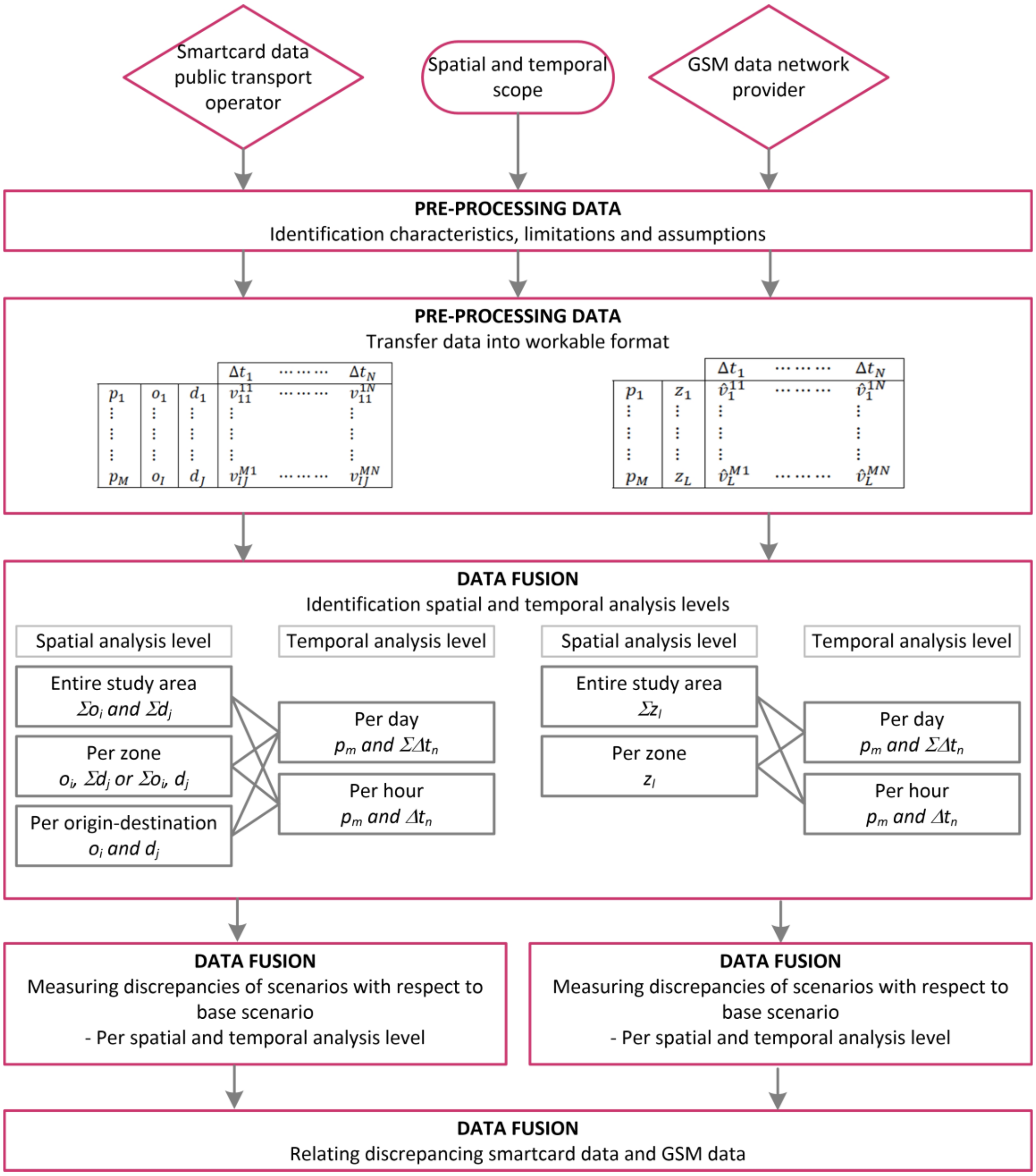

FIGURE 1 Workflow on the data analysis process

2.1 Pre-processing data

The smartcard and the GSM datasets have different characteristics and limitations. These are first described per dataset before turning into the data fusion. For more information concerning the data formats, the reader is referred to *(19).*

2.1.1 Smart card data

The smart card data used for this research are anonymous OV-chipkaart data. In the Netherlands, the OV-chipkaart is used nationwide for public transport fare validation. All passengers have to tap in and tap out. Each smartcard transaction record contains information on the origin, i, and destination, j, at the stop level and the respective time stamp. Transactions are then temporally aggregated per day, m, and time intervals, n. The aggregations results with passenger volume denoted by v_{ij}^{mn} travelling from origin i to destination j, on a specific day m, during time interval n.

2.1.2 GSM data

GSM data for this study were provided by DAT.Mobility, who in turn receives data from a network provider (Vodafone) with a market share of approximately 33% in the Netherlands. The data received are already completely anonymized such that individuals cannot be traced *(16).* The data reports the amount of devices counted per spatial and temporal features for all Vodafone users, and a growth factor algorithm is applied to increase the sample to the total population. The resulting data have been validated by DAT.Mobility and Bureau of Statistics in the Netherlands, and its accuracy has been verified *(20).*

Each time a phone connects to the network, it is detected and registered in the database. A telephone that is switched on, connects approximately 20 times to a network per day, even if it is not actively used. An actively used device connects more often to the network. Based on the antenna the device connects to, the location of the device is estimated. Antennas, however, cover multiple areas and multiple antennas may cover the same area *(14).* As a result, there is a localization error when estimating the location of the device *(13).* To ensure a high level of accuracy of the spatial features in the GSM data, zones are defined, to which devices are allocated. The zones included in the GSM data, cover a larger geographical area than the catchment area of stop-level smartcard data. The geographical size of the zones may strongly vary, based on one or more postal codes areas in the Netherlands; i.e. zones of 6 up to 30 km^2 are found.

The GSM data available for this study was temporally aggregated into pre-defined time periods. The allocation algorithm searches for unique devices per time interval. If a device is detected in multiple zones within a single time interval, it is allocated to the zone in which it has been detected for the longest period of time within the respective period. Furthermore, a distinction is made between visitors and residents. To determine whether a device belongs to a visitor or a resident of that zone, the place of residence of each device is estimated based on overnight detections. The zone in which the device is detected in the majority of the nights during one month, is determined to be the place of residence of that device. The process is performed each month, since the data are monthly encrypted. If a device is detected in its place of residence, it is registered as a resident, otherwise it is registered as a visitor. Due to the spatial aggregation, it is not possible to determine whether a device stayed at home or moved within the zone when recorded in its zone of residency. In contrast, visitors moved from their place of residence to another zone, thereby manifesting demand for transport. Given the purpose of this study, only visitors were included in further analysis.

The GSM occupancy data contains information concerning $\hat{v}_l^{mn}$, the number of visitors detected in zone $l \in L$ during day m and time interval n. L is the set of zones defined in the case study area. The place of residence is not included; hence, it is unknown where visitors come from. Furthermore, the difference between two subsequent hours is a net change in zone occupancy: the arrival-departure ratio cannot be deduced. Demand for transport is investigated using the net change of visitors, the absolute level of demand for transport cannot be deduced.

2.2 Data fusion

2.2.1 Spatial and temporal analysis levels

To ensure consistency, the smartcard data are aggregated accordingly: for each zone, transactions recorded at stops within a certain time interval are summed. By aggregating and differentiating spatial and temporal features of the datasets, different analysis levels are identified, for which scenarios can be analyzed. Spatial analysis is performed for the entire study area, per zone or per origin-destination relation. The latter is possible only for the smart card data and not for the fused data. Temporal analysis is performed at the hourly and daily levels. Intersecting the spatial and temporal analysis levels leads to four combinations: total daily, total hourly, zonal daily and zonal hourly. The total daily level hereby gives a high-level overview of the data, whereas each of the following levels zooms into spatial, temporal or both features. This top-down approach is commonly used to analyze (public) transport mobility patterns *(16,21,22).*

2.2.2 Measuring discrepancies per dataset

For each dataset and analysis level the discrepancies are measured in comparison to the respective base scenario. Normalized discrepancies are measured, in order to allow the comparison of results obtained for two different data sources. In addition, the direction and magnitude of the discrepancies should be considered. We therefore chose to use the Mean Percentage Error (MPE) measure. The formulas differ per analysis level and the values and features included in the dataset under consideration (v_{ij}^{mn} for the smartcard data and $\hat{v}_{l}^{mn}$ for the GSM data). For the smartcard data, in the zonal hourly analysis level, a distinction can be made between arrivals or departures per zone. Eq. (1)-(3) provide the MPE definitions for the smartcard data and Eq. (4)-(5) define the MPE for the GSM data.

$$Total\ hourly\ MPE_{smartcard,n} = \frac{1}{I \cdot J} \cdot \frac{\left(\sum_i \sum_j v_{ij}^{[scenario]n} - \sum_i \sum_j v_{ij}^{[base]n}\right)}{\sum_i \sum_j v_{ij}^{[base]n}} \quad (1)$$

$$Zonal\ hourly\ MPE_{smartcard,jn} = \frac{1}{I} \cdot \frac{\left(\sum_i v_{ij}^{[scenario]n} - \sum_i v_{ij}^{[base]n}\right)}{\sum_i v_{ij}^{[base]n}} \quad (2)$$

$$Zonal\ hourly\ MPE_{smartcard,in} = \frac{1}{J} \cdot \frac{\left(\sum_j v_{ij}^{[scenario]n} - \sum_j v_{ij}^{[base]n}\right)}{\sum_j v_{ij}^{[base]n}} \quad (3)$$

$$Total\ hourly\ MPE_{GSM,n} = \frac{1}{|L|} \cdot \frac{\left(\sum_l \hat{v}_{l}^{[scenario]n} - \sum_l \hat{v}_{l}^{[base]n}\right)}{\sum_l \hat{v}_{l}^{[base]n}}$$

(4)

$$Zonal\ hourly\ MPE_{GSM,ln} = \frac{\left(\hat{v}_{l}^{[scenario]n} - \hat{v}_{l}^{[base]n}\right)}{\hat{v}_{l}^{[base]n}} \quad (5)$$

The MPE values are in the range $[-1, \infty)$. In the following, if the MPE falls within the user-defined range $[-0.2, 0.2]$ then the respective analysis unit is considered not significantly different from the base scenario.

2.2.3 Relating discrepancies of smart card data and GSM data

The final step in the data fusion procedure is relating the smartcard metrics with the GSM metrics. The relation between MPE values is established by means of a graph, plotting the MPE values of both datasets on the axes, as illustrated in Figure 2. The threshold value range is displayed using pink dotted lines. If the dots follow the grey dotted line, this means the relative MPE values of the public transport usage are of the same order as the relative MPE values of the visitor occupancy. The non-shades areas in the graph are of most interest for public transport operators. For example, in the time interval 11:00-12:00 the visitor occupancy increased significantly compared to the base scenario, whereas the public transport usage significantly decreased relatively to the base scenario. It is highly relevant for the public transport operator to examine why the public transport usage falls while general demand for transport increases for this area and time period.

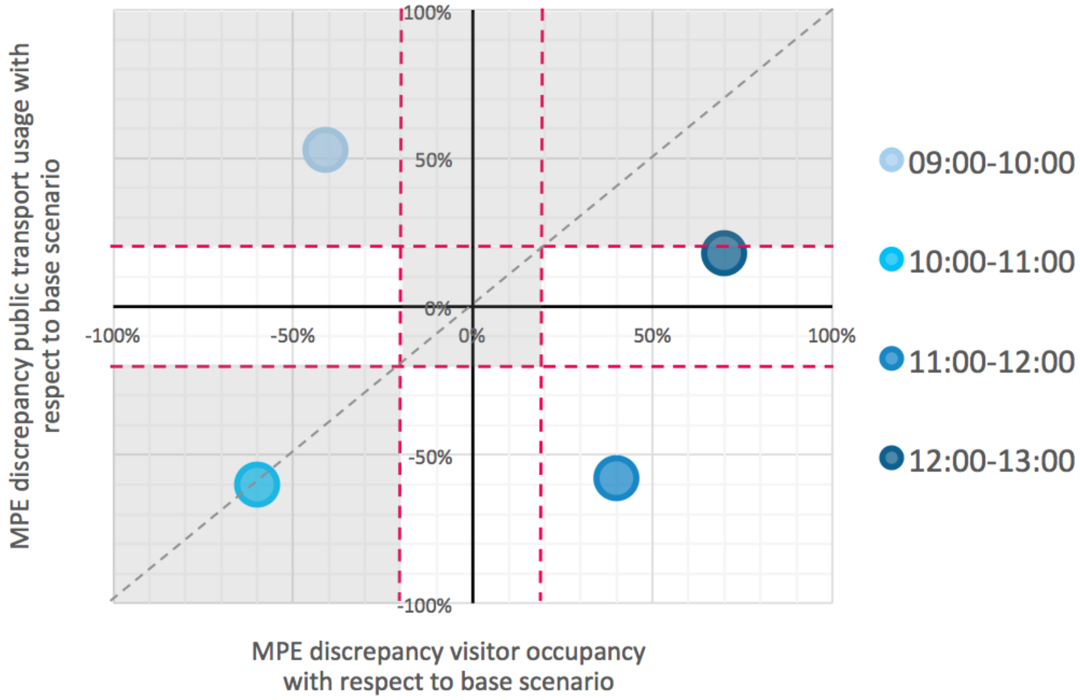

FIGURE 2 An illustration of relating Mean Percentage Error of public transport usage and visitor occupancy for a given area and time periods when compared to the base level

3. CASE STUDY: LATE EVENINGS AND EARLY MORNINGS IN ROTTERDAM

3.1 Case study description

We applied our methodology to two case studies: (a) special events (e.g. festivals, disturbances) in Amsterdam and their respective mobility and transit patterns; (b) night service in Rotterdam. Only the latter is presented here due to space limitations. The details of the Amsterdam case study are available in *(19)*.

Rotterdam is the second largest city in the Netherlands, with approximately 600,000 inhabitants. RET is the public transport operator in the city and surroundings, operating bus, tram and metro services. On yearly basis, approximately 160 million passenger trips are performed with RET *(24)*. The case study area includes 34 zones, based on the availability of urban public transport network throughout the late evenings and early mornings (Figure 3).

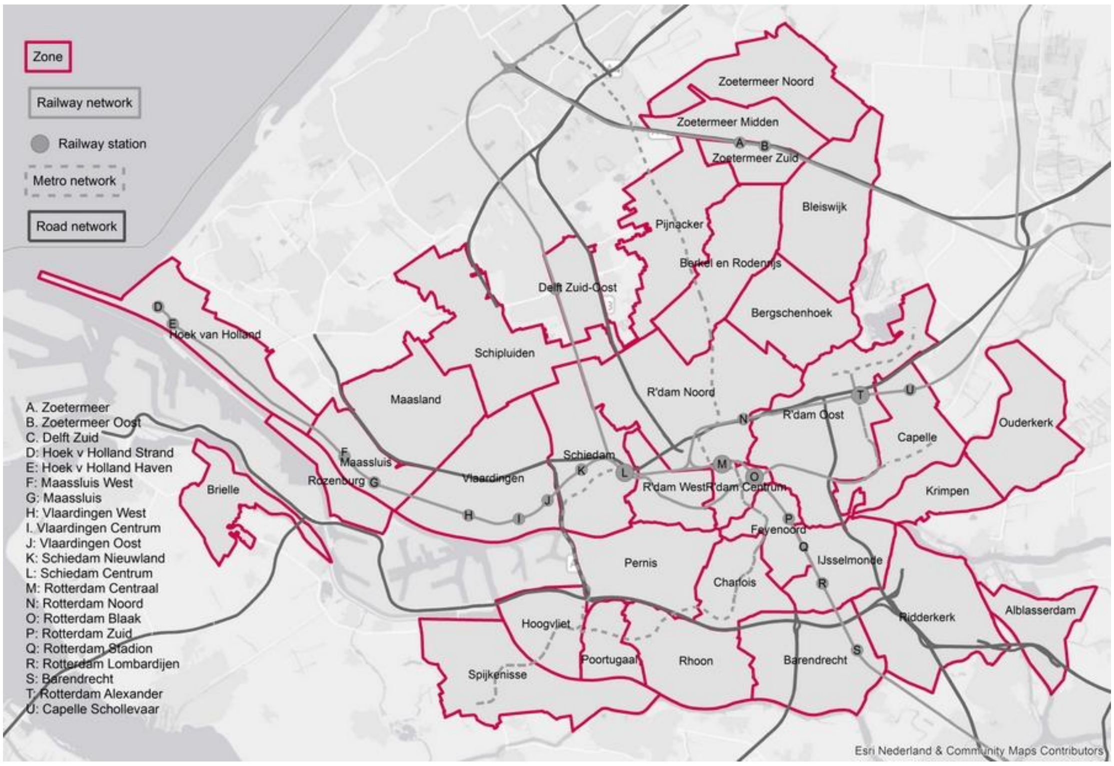

FIGURE 3 Spatial demarcation of the Rotterdam case study area

The case study was designed to analyze whether the service span of the public transport network is in line with a respective increase and decrease in the overall travel demand. For example, it may be shown that according to the overall transport demand it is useful for a specific type of day to extend the public transport operations in the late evenings, or to start operating earlier in the morning. All working days from January 5^{th} to May 31^{th} in 2015 are taken into account with the exception of few days were large-scale events took place. The starting and ending time of the transit operations may differ per zone. Operations end between midnight and 2AM, whereas operations are resumed again between 5AM and 7AM. In total, 84 nights are included in the analysis. The results are reported based on the average mobility patterns observed from the smartcard and the GSM data. The results are presented with respect to the relative change in comparison to the previous hour. In case of visitor occupancy as measured by GSM data, a decrease with respect to the previous hour shows demand for outbound transport from a given zone, whereas an increase indicates demand for inbound transport towards the zone. The results for the total and zonal hourly analysis levels are presented in the following sub-sections.

3.2 Total hourly analysis results

The total hourly MPE values on working days are displayed in Figure 4. It can be observed that in the late evening hours and until 2AM, both visitor occupancy and public transport usage decrease on an hour-on-hour basis with the latter decreasing much more sharply than the former. This may suggest that the service ends too early given that reductions in visitors level exhibit a slower pace using travel modes other than public transport. During the night, between 2AM and 5AM, no significant changes are observed, whereas in the early mornings, from 5AM onwards, a rapid increase in both public transport usage and visitor occupancy is measured. At the aggregate level, the transition from night to daytime network seem to be justified given the simultaneous change in overall occupancy levels.

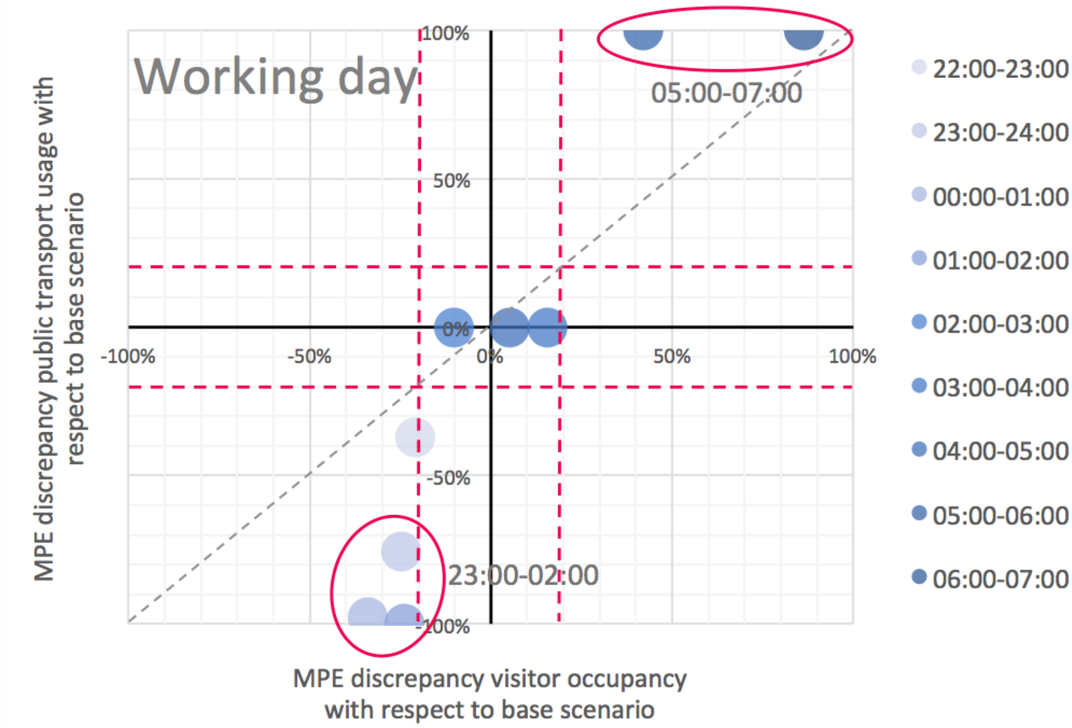

FIGURE 4 Discrepancies of public transport usage and visitor occupancy (hourly; working day) related to the base scenario

3.3 Zonal hourly analysis results

Investigating hourly changes at the zone level allow identifying where the night service might be inadequate. The results indicate that visitor occupancy continues to change in several zones in the late evening after public transport services ceased (1AM-2AM) and increase considerably in the early morning when they are gradually resumed (5AM-6AM). The results for these two time intervals are visualized in Figure 5. The background color of each zone shows the relative MPE value of visitor occupancy with respect to the previous hour, and the color of the circle within each zone shows the corresponding value for public transport usage. If no circle is included in the zone, no public transport data are available for the working days. Minimum MPE values for the smartcard data are -100%, whereas this is -37% for the GSM data. Maximum MPE values for the smartcard follow from the usage in the first operating hour, as an increase to no operations the hour before. Maximum MPE value for GSM is +88%. Light red and light green imply a decrease or increase within the threshold value range; i.e. the range defined as a non-significant change (section 2.2.2). Zones with contradictory colors are of particular interest. For the time interval 1AM-2AM (Figure 5, top part) zones of interest are found especially in the Northern and Southern parts of the case study area (dark red background, light red circle). These suburban and residential zones have a significant decrease in visitor occupancy during this hour, which cannot be served by public transport since their operations have already stopped. For the time interval 5AM-6AM (Figure 5, bottom part) zones of interest are found especially in the Western and Southern part of the case study area (dark green background, light red dot). These industrial and logistic zones around the large port area already have a significant increase in visitor occupancy during this hour with respect to the previous hour, whereas public transport operations have not resumed yet.

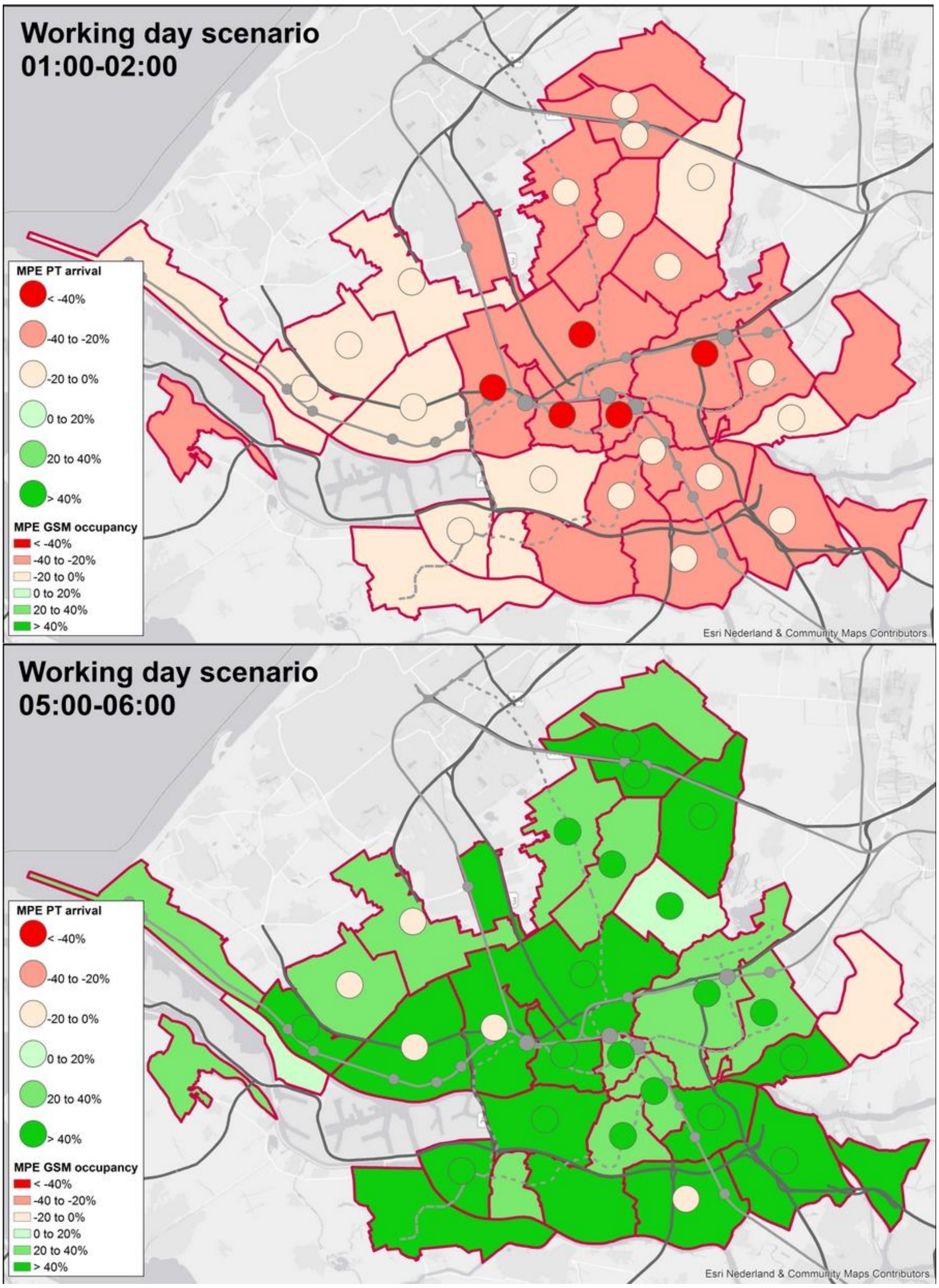

FIGURE 5 Mean Percentage Error hour-on-hour values on working days for visitor occupancy (background color) and public transport usage (circle color) for time intervals 1AM-2AM (top) and 5AM-6AM (bottom)

The MPE calculates relative changes in order to allow relating the two data sources to each other. The net absolute change in visitor occupancy compared with the previous hour is however also of interest for the local operator in order to assess the magnitude of the potential demand. Tables 1 and 2 summarize the relative change of both the public transport usage and the visitor occupancy for time intervals 1AM-2AM and 5AM-6AM compared with the previous hour for the zones of interest, identified based on Figure 5. The value of the net change in visitor occupancy is also given. The zones directly south of the Maas river, Feyenoord and Ridderkerk (Table 1, Figure 3), where much of the nightlife activities are concentrated, see a substantial decrease of at least 1,000 people during the late night hours.

This is the lower limit of the number of people that change their location during this hour since the change in occupancy corresponds to the net change, indicating therefore for a potential for public transport services during these hours. It is especially important to cater for this demand due to alcohol consumption that is customary for nightlife.

During the early morning, between 5Am and 6AM a net change of 1,600 in visitor occupancy is observed in Barendrecht (Table 2, Figure 3), a factory area, hence a large inbound demand can be targeted by the operator. In contrast, it can be concluded that Maasland and Schipluiden are not much of interest for the operator given the low absolute changes in visitor occupancy. For the other zones, i.e. Schiedam and Vlaardingen, a relatively high absolute value of visitor occupancy is observed suggesting that there is a potential demand for additional public transport in the early mornings.

TABLE 1 Relative and absolute changes in visitor occupancy for selected zones during time interval 1AM-2AM on working days

Zone name	MPE of visitor occupancy	Net change in visitor occupancy
Barendrecht	-27%	550
Bergschenhoek	-23%	200
Berkel & Rodenrijs	-27%	300
Feyenoord	-27%	1,100
IJsselmonde	-23%	550
Pijnacker	-24%	150
Ridderkerk	-37%	1,000
Zoetermeer Midden	-25%	500
Zoetermeer Zuid	-29%	300

TABLE 2 Relative and absolute changes in visitor occupancy for selected zones during time interval 5AM-6AM on working days

Zone name	MPE of visitor occupancy	Net change in visitor occupancy
Barendrecht	+88%	1,600
Maasland	+33%	180
Schiedam	+60%	1,300
Schipluiden	+35%	75
Vlaardingen	+51%	1,000

4. CONCLUSIONS AND RECOMMENDATIONS

The public transport industry faces challenges to cater for the variety of mobility patterns and corresponding needs and preferences of passengers. Even though data fusion can potentially be used to investigate spatial and temporal variations in travel demand, it is only seldom used by public transport operators. We developed a methodology to fuse smartcard and GSM data to allow analyzing public transport usage in relation to the overall travel demand. Based on the relation of relative changes in public transport usage and visitor occupancy for different analysis levels, spatial and temporal features of interest for public transport operators can be examined. The analysis approach proposed in this study supports public transport operator decision making at the tactical level.

Due to different semantics of the smartcard and GSM data, it is not possible to directly fuse both datasets. Our methodology, however, demonstrated the systematic exploration and analysis of public transport usage in relation to the overall travel demand. This information could not be deduced by analyzing a single dataset. Due to the spatial level of detail of the GSM data, it is not possible to determine exact locations of demand for transport, and origin-destination relations are unknown. However, the application of the methodology to a case study in the Netherlands, showed the identification of several zones that are of interest for the public transit operator; i.e. showing a potential demand for extending the service span both in the late evening and early morning. The potential demand for public transport in turn has to be considered in more detail, while taking into account the possible line alignments and public transport market share, since not all the mobility change will shift in response to service provision. In addition, in order to identify whether it would be useful to extend public transport operations beyond the current service span, capacity utilization and cost estimates are needed.

The data fusion approach proposed in this paper can be used to explore and fuse a large range of datasets that contains information (in aggregated or disaggregated form) for origins and/or destinations in transport networks. Several limitations of the methodology can be identified, pertaining to data processing issues. Even though ongoing efforts decrease the size of the zones used in the aggregation of the GSM data in the Netherlands, privacy concerns dictate that considerable aggregation will remain *(13).* For future improvements of the methodology, the inclusion of origin-destination relations in the GSM data would provide information on the direction of the potential public transport demand. Smartcard data in the Netherlands is owned and stored by individual public transport operators. Fusing data from different operators, including the national railway, will enable identifying passengers transferring between services provided by different operators.

ACKNOWLEDGMENTS

This research is performed in cooperation with Goudappel Coffeng, DAT.Mobility, RET and Delft University of Technology, Department of Transport & Planning.

REFERENCES

1. Guedes, M. C. M., Oliveira, N., Santiago, S., & Smirnov, G. On the evaluation of a public transportation network quality: Criteria validation methodology. *Research in Transportation Economics*, Vol. 36(1), 2012, pp. 39-44.
2. Cats, O., Wang, Q., & Zhao, Y. Identification and classification of public transport activity centres in Stockholm using passenger flows data. *Journal of Transport Geography*, Vol. 48, 2015, pp. 10-22.
3. Gutierrez, J., & García-Palomares, J. C. New spatial patterns of mobility within the metropolitan area of Madrid: towards more complex and dispersed flow networks. *Journal of transport geography*, Vol. 15(1), 2007, pp. 18-30.
4. Pelletier, M. P., Trépanier, M., & Morency, C. Smart card data use in public transit: A literature review. *Transportation Research Part C: Emerging Technologies*, Vol. 19(4), 2011, pp. 557-568.
5. Elias, D., Nadler, F., Stehno, J., Krösche, J., & Lindorfer, M. SOMOBIL–Improving Public Transport Planning Through Mobile Phone Data Analysis. *Transportation Research Procedia*, Vol. 14, 2016, pp. 4478-4485.
6. Durand, C. P., Tang, X., Gabriel, K. P., Sener, I. N., Oluyomi, A. O., Knell, G., Porter, A.K., Hoelscher, D. M. & Kohl, H. W. The association of trip distance with walking to reach public transit: Data from the California Household Travel Survey. *Journal of Transport & Health*, Vol. 3(2), 2016, pp. 154-160.
7. Long, Y., & Thill, J. C. Combining smart card data and household travel survey to analyze jobs–housing relationships in Beijing. *Computers, Environment and Urban Systems*, Vol. 53, 2015, pp. 19-35.
8. Del Castillo, J. M., & Benitez, F. G. A methodology for modeling and identifying users satisfaction issues in public transport systems based on users surveys. *Procedia-Social and Behavioral Sciences*, Vol. 54, 2012, pp. 1104-1114.
9. Frias-Martinez, V., Soguero, C. and Frias-Martinez, E. Estimation of urban commuting patterns using cellphone network data. In *Proceedings of 6th Transport Research Arena, April 18-21, 2016, Warsaw, Poland,* 2016.

10. Van Oort, N., Sparing, D., Brands, T. and Goverde, R. M. P. Data driven improvements in public transport: the Dutch example. *Public Transport,* Vol. 7(3), pp. 369-389.
11. Van der Mede, P. Over het meten van mobiliteit met GSM-data: mogelijkheden en onmogelijkheden. Contribution for *Colloquium Vervoersplanologisch Speurwerk, Eindhoven, 2014. (in Dutch)*
12. Aguilera, V., Allio, S., Benezech, V., Combes, F. and Milion, C. Using cell phone data to measure quality of service and passenger flows of Paris transit system. *Transportation Research Part C: Emerging Technologies*, Vol. 43(2), 2014, pp. 198-211.
13. Calabrese, F., Diao, M., Di Lorenzo, G., Ferreira, J., & Ratti, C. Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. *Transportation research part C: emerging technologies*, Vol 26, 2013, pp. 301-313.
14. Iqbal, M. S., Choudhury, C. F., Wang, P., & González, M. C. Development of origin–destination matrices using mobile phone call data. *Transportation Research Part C: Emerging Technologies*, Vol. 40, 2014, pp. 63-74.
15. Kusakabe, T. and Asakura, Y. Behavioural data mining of transit smart card data: A data fusion approach. *Transportation Research Part C: Emerging Technologies*, Vol. 46, 2014, pp. 179-191.
16. Elfrink, M., Courtz, M., Metz, S., Ebben, M., and Weppner, J. OV-potentie opsporen door datafusie. *Nationaal Verkeerskundecongres*, 2015. *(in Dutch)*
17. Holleczek, T., Yu, L., Lee, J. K., Senn, O., Ratti, C., & Jaillet, P. Detecting weak public transport connections from cellphone and public transport data. In *Proceedings of the 2014 International Conference on Big Data Science and Computing* (p. 9). ACM, 2014.
18. Duff-Riddell, W. R. and Bester, C. J. Network modeling approach to transit network design. *Journal of Urban Planning and Development*, Vol. 131(2), 2005, pp. 87-97.
19. De Regt, K.L. How do spatial and temporal patterns of public transport relate to the overall travel demand? A data fusion method for smart card data and GSM data, *Master thesis, Delft University of Technology*, 2016.
20. ViewDAT. *ViewDAT*, http://view.dat.nl/viewdat/. Accessed October 2015.
21. Liu, L., Hou, A., Biderman, A., Ratti, C. and Chen, J. Understanding individual and collective mobility patterns from smart card records: A case study in Shenzhen. In *Intelligent Transportation Systems, 2009. ITSC' 09. 12th International IEEE Conference*, p. 1-6, 2009.
22. Nishiuchi, H., King, J. and Todoroki, T. Spatial-temporal daily frequent trip pattern of public transport passengers using smart card data. *International Journal of Intelligent Transportation Systems Research*, Vol. 11(1), 2013, pp. 1-10.
23. RET, *Over RET*, http://corporate.ret.nl. Accessed July 2016. *(in Dutch)*