Investigating Potential Transit Ridership by Fusing Smartcard Data and GSM Data

de Regt, Karin; Cats, Oded; van Oort, Niels; van Lint, Hans

DOI
10.3141/2652-06

Publication date
2017

Document Version
Final published version

Published in
Transportation Research Record

Investigating Potential Transit Ridership by Fusing Smartcard and Global System for Mobile Communications Data

Karin de Regt, Oded Cats, Niels Van Oort, and Hans van Lint

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

Both passengers and the government demand an efficient public transport system that is of both high quality and low cost. This system has to be user-oriented, and it must 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 public transport travel demand, neglecting overall travel demand, although this 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 household travel survey data is a time-intensive and costly undertaking, primarily because of the labor-intensive process of acquiring and processing surveys. As a result, such surveys are performed at long intervals (measured in years), aiming to represent an average (working) day for travelers (*9*). It is therefore not possible to distinguish temporal dynamics, because only an average day is represented. This calls for the development of methods designed to acquire information on both the spatial and temporal dynamic mobility patterns of public transport passengers in relation to overall travel demand.

In addition to smartcard data and travel habit surveys, several other data sources have been used to gain information on mobility patterns and improve the design of public transport. Examples of these data sources include automatic vehicle location systems, Wi-Fi and Bluetooth signals, social media, and the Global System for Mobile Communications (GSM) (*10*). The most important challenge is to process the data so that they become useful for improving public transport design. Although automatic vehicle location systems allow the monitoring of fleet performance, they do 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 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, vary considerably from country to country. GSM data are extracted from call-detail records that are supplied by the network provider (*11*). Three main applications of GSM data in transport research are origin–destination estimation, detection of events on the basis of crowdedness, and travel mode identification (*12–14*). The latter is not yet applicable for GSM data in the Netherlands; therefore, relying solely on GSM data was not sufficient for the purpose of this study.

The combination of data sources known as data fusion offers a promising avenue for gaining information on public transport mobility patterns versus overall temporal and spatial travel demand. Several data fusion studies considered either smartcard data or GSM data combined with travel habit surveys to successfully estimate trip purposes (*7, 15*). A pilot data fusion study was performed in Emmen, the Netherlands, where smartcard and GSM data were fused to find areas with the potential to support additional public transport (*16*). A study in Singapore also explored the combination of smartcard and GSM data to identify weak public transport connections (*17*). Both studies supported the hypothesis that the data fusion of smartcard and GSM data offers synergies resulting with new information (*16*). Smartcard

K. de Regt, Transport, Infrastructure, and Logistics, and O. Cats, N. van Oort, and H. van Lint, Transport and Planning, Faculty of Civil Engineering and Geosciences, Delft University of Technology, P.O. Box 5048, 2600 GA Delft, Netherlands. Corresponding author: O. Cats, O.Cats@tudelft.nl.

Transportation Research Record: Journal of the Transportation Research Board, No. 2652, 2017, pp. 50–58.
http://dx.doi.org/10.3141/2652-06

data provide information on the public transport passengers traveling with a specific operator. GSM data provide information on overall travel demand and its temporal and spatial variation. Although all transport modalities are included, no distinction can be made between the different modalities. Both data sources contain information on spatial and temporal variations.

The objective of this study was to analyze the potential of fusing smartcard and GSM data for gaining information on public transport mobility patterns, versus 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 so that the network design and service span live up to the demand for transport during the night. The data fusion analysis approach was applied to a case study in Rotterdam, Netherlands, in which public transport usage versus overall travel demand was analyzed for the late evenings and early mornings for different types of days. The aim was to identify whether the supply of urban public transport services was adequate for the demand for transport during these hours, and in what way this varies on different types of days. The results of this study can support decision makers in evaluating a service design and schedule and identify potential improvements.

The next section explains the proposed data fusion methodology, followed by a description of the application of the methodology to the case study of the night services in Rotterdam. Then the findings and recommendations for future applications are presented.

METHODOLOGY: FUSION GSM AND SMARTCARD DATA

In this 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 was anonymized smartcard and GSM data along with the relevant spatial and temporal information. Depending on the application of interest, a base case scenario was defined (e.g., representing conditions on an average day or referring to a moving reference level, such as the previous hour). Input data preprocessing consisted of two aspects: identification of the characteristics, limitations, and assumptions of each data set and processing the data into a workable format. After preprocessing, the data fusion could be performed. Here, first different spatial and temporal analysis levels were identified by means of aggregation or differentiation in space and time. By data set and by analysis level, the discrepancies of scenarios with respect to the base scenario were measured using quantitative metrics. The actual data fusion was established by relating the discrepancies of the smartcard data to the discrepancies of the GSM data by scenario and analysis level. The approach proposed in this study can be used to explore various data sets that contains information on origins and destinations, or both, in transport networks.

Preprocessing Data

The smartcard data set and the GSM data set have different characteristics and limitations. These are first described by data set before turning to the data fusion. More information on the data formats can be found in the work of de Regt (*19*).

Smartcard Data

The smartcard data used for this research were 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 contained information on the origin, *i,* and destination, *j,* at the stop level, and on the time stamp. Transactions were then temporally aggregated per day, *m,* and time intervals, *n.* The aggregation results showed passenger volume, denoted by v_{ij}^{mn}, traveling from origin *i* to destination *j,* on a specific day *m,* during time interval *n.*

GSM Data

GSM data for this study were provided by DAT.Mobility, which in turn receives data from a network provider (Vodafone) with a market share of approximately 33% in the Netherlands. The data received were already completely anonymized; individuals could not be traced (*16*). The data reported the number of devices counted per spatial and temporal features for all Vodafone users, and a growth factor algorithm was applied to increase the sample to the total population. The resulting data were validated by DAT.Mobility and the Bureau of Statistics in the Netherlands, and the 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. On the basis of 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 were defined and devices were allocated to those zones. The zones included in the GSM data covered a larger geographical area than the catchment area of stop-level smartcard data. The geographical size of the zones may vary considerably, on the basis of one or more postal code areas in the Netherlands; that is, zones of 6 km^2 up to 30 km^2 were found.

The GSM data were provided by one of the largest mobile phone operators in the Netherlands. The very large sample was then scaled by applying a growth factor algorithm. The algorithm used national census data while accounting for seasonality, type of day, time of day, and average duration of activities. Accuracy verification was performed by comparing the maximum amount of inhabitants per zone resulting from the growth algorithm with the zone population data available from the Central Bureau of Statistics. The algorithm was then adjusted to attain population levels that were very similar to the official statistics. Similar methods of scaling GSM data to the total population have been applied in other transport studies (*5, 17, 21*).

The GSM data available for this study were temporally aggregated into a set of time periods predefined by DAT.Mobility, the company that owns and manages this data set. The allocation algorithm searched for unique devices per time interval. If a device was detected in multiple zones within a single time interval, it was allocated to the zone in which it had been detected for the longest period

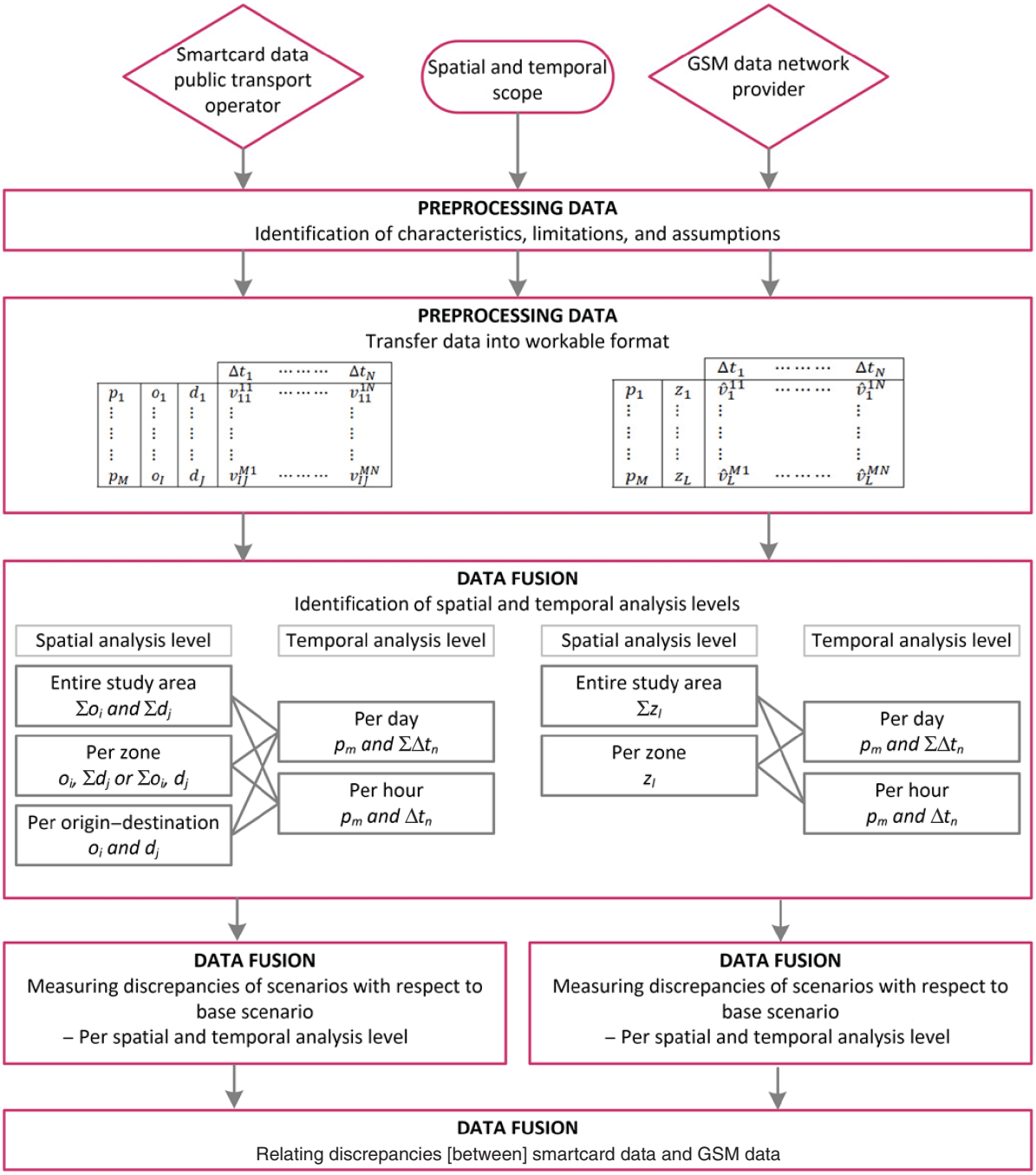

FIGURE 1 Workflow of the data analysis process.

of time within the respective period. Furthermore, a distinction was made between visitors and residents. To determine whether a device belonged to a visitor or a resident of that zone, the place of residence of each device was estimated on the basis of overnight detections. The zone in which the device was detected in most nights during one month was determined to be the place of residence of that device. The process was performed each month, because the data are encrypted monthly. If a device was detected in its place of residence, it was registered as a resident; otherwise it was registered as a visitor. Because of the spatial aggregation, it was not possible to determine whether a device stayed at home or moved within the zone when it was 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 contained information concerning the number of visitors detected in zone $l \in L$ during day m and time interval n, denoted by $\hat{v}_l^{mn}$. L is the set of zones defined in the case study area. The place of residence was not included; hence, it is unknown where visitors came 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 was investigated by using the net change in number of visitors; the absolute level of demand for transport cannot be deduced.

Data Fusion

Spatial and Temporal Analysis Levels

To ensure consistency, the smartcard data were aggregated by zone: for each zone, transactions recorded at stops within a certain time interval were summed. By aggregating and differentiating spatial and temporal features of the data sets, different analysis levels were identified for which scenarios can be analyzed. Spatial analysis was performed for the entire study area, per zone or per origin–destination relation. The latter was possible only for the smartcard data and not for the fused data. Temporal analysis was performed at the hourly and daily levels. Intersecting the spatial and temporal analysis levels led to four combinations: total daily, total hourly, zonal daily, and zonal hourly. The total daily level gives a high-level overview of the data; the other levels zoom into spatial features, temporal features, or both. This top-down approach is commonly used to analyze (public) transport mobility patterns (*16, 22, 23*).

Measuring Discrepancies per Data Set

For each data set and analysis level, the discrepancies were measured in comparison with the respective base scenario. Normalized discrepancies were measured to allow the comparison of results obtained for two different data sources. In addition, the direction and magnitude of the discrepancies were considered. The mean percentage error (MPE) measure was therefore chosen. The formulas differed by analysis level and by the values and features included in the data set 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 could be made between arrivals or departures per zone. Equations 1 through 3 provide the MPE definitions for the smartcard data and Equations 4 and 5 define the MPE for the GSM data.

$$\text{total hourly MPE}_{\text{smartcard},n} = \frac{1}{I \cdot J} \cdot \frac{\left(\sum_i \sum_j v_{ij}^{[\text{scenario}]n} - \sum_i \sum_j v_{ij}^{[\text{base}]n} \right)}{\sum_i \sum_j v_{ij}^{[\text{base}]n}} \tag{1}$$

$$\text{zonal hourly MPE}_{\text{smartcard},j,n} = \frac{1}{I} \cdot \frac{\left(\sum_i v_{ij}^{[\text{scenario}]n} - \sum_i v_{ij}^{[\text{base}]n} \right)}{\sum_i v_{ij}^{[\text{base}]n}} \tag{2}$$

$$\text{zonal hourly MPE}_{\text{smartcard},in} = \frac{1}{J} \cdot \frac{\left(\sum_j v_{ij}^{[\text{scenario}]n} - \sum_j v_{ij}^{[\text{base}]n} \right)}{\sum_j v_{ij}^{[\text{base}]n}} \tag{3}$$

$$\text{total hourly MPE}_{\text{GSM},n} = \frac{1}{|L|} \cdot \frac{\left(\sum_l \hat{v}_l^{[\text{scenario}]n} - \sum_l \hat{v}_l^{[\text{base}]n} \right)}{\sum_l \hat{v}_l^{[\text{base}]n}} \tag{4}$$

$$\text{zonal hourly MPE}_{\text{GSM,ln}} = \frac{\left(\hat{v}_l^{[\text{scenario}]n} - \hat{v}_l^{[\text{base}]n} \right)}{\hat{v}_l^{[\text{base}]n}} \tag{5}$$

The MPE values are in the range [−1, ∞). If the MPE fell within the user-defined range [−0.2, 0.2], then the respective analysis unit was considered not significantly different from the base scenario.

Relating Discrepancies of Smartcard Data and GSM Data

The final step in the data fusion procedure was relating the smartcard metrics with the GSM metrics. The relation between MPE values was established by means of a graph, plotting the MPE values of both data sets on the axes (Figure 2). The threshold value range is displayed by using pink dotted lines. If the dots follow the gray dotted line, this means the relative MPE values of public transport usage are of the same order as the relative MPE values of visitor occupancy. The unshaded areas in the graph are of most interest for public transport operators. For example, in the time interval 11:00 a.m. to noon, the visitor occupancy increased significantly compared with the base scenario, whereas the public transport usage significantly decreased relative to the base scenario. It is highly relevant for the public transport operator to examine why the public transport

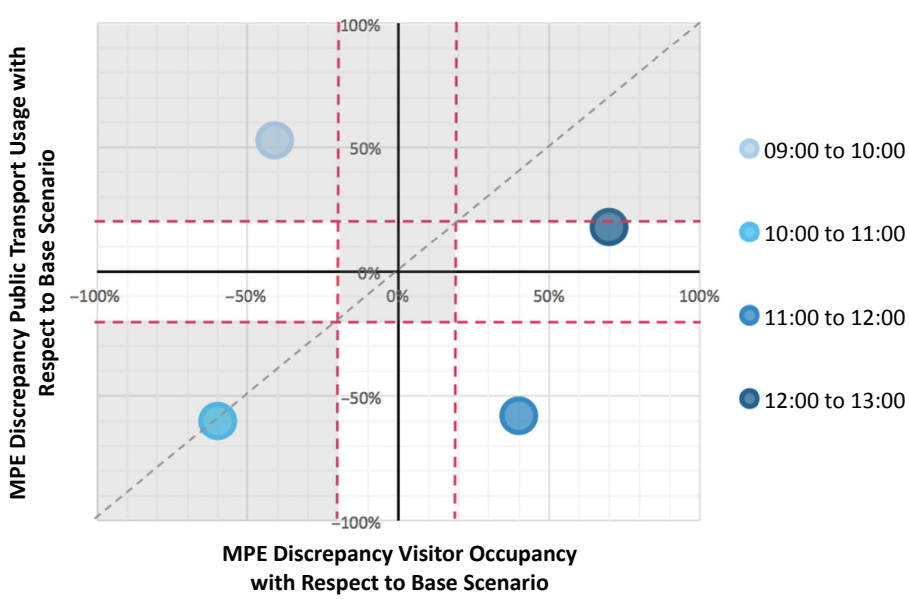

FIGURE 2 MPE of public transport usage and visitor occupancy compared with the base level.

usage falls while the general demand for transport increases for this area and time period.

CASE STUDY: LATE EVENINGS AND EARLY MORNINGS

Case Study Description

The methodology was applied to two case studies: (*a*) special events (e.g., festivals, disturbances) in Amsterdam and their respective mobility and transit patterns and (*b*) night service in Rotterdam. Only the latter is presented here because of space limitations. The details of the Amsterdam case study are available in de Regt (*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 a yearly basis, approximately 160 million passenger trips are made with RET (*24*). The case study area includes 34 zones, on the basis of the availability of the urban public transport network throughout the late evenings and early mornings (Figure 3).

The case study was designed to analyze whether the service span of the public transport network was in line with increases and decreases in the overall travel demand. For example, it may be shown that, according to the overall transport demand, it is useful on a specific type of day to extend public transport operations in the late evening, or to start operating earlier in the morning. The starting and ending time of the transit operations may differ per zone. Operations end between midnight and 2:00 a.m. and are resumed again between 5:00 and 7:00 a.m.

Following the preprocessing of the smartcard and GSM data, both data sets were processed for the specific spatial and temporal demarcations of this case study. The smartcard data were aggregated from stop level to the level of the zones for which data were available from the GSM data set. All working days from January 5 to May 31 in 2015 were taken into account, with the exception of a few days on which large-scale events took place. In total, 84 nights (10:00 p.m. to 7:00 a.m.) were included in the analysis. Data sets were processed and analyzed in MATLAB and ArcGIS, including its Python toolboxes.

The results are reported on the basis of the average mobility patterns observed from the smartcard and GSM data. The results are presented with respect to the relative change in comparison with the previous hour. In case of visitor occupancy, as measured by the GSM data, a decrease with respect to the previous hour showed demand for outbound transport from a given zone, whereas an increase indicated demand for inbound transport toward the zone.

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 2:00 a.m., both visitor occupancy and public transport usage

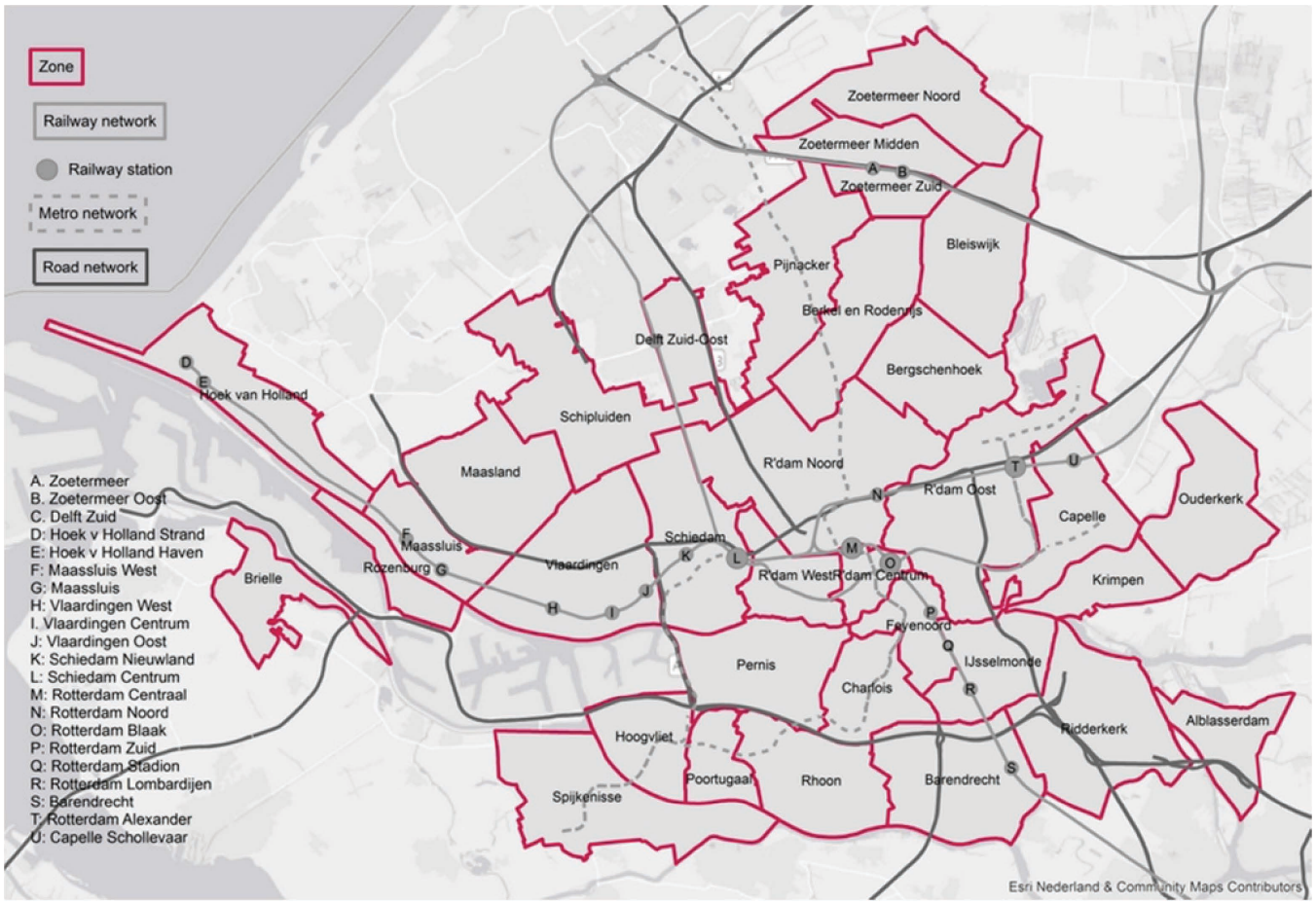

FIGURE 3 Spatial demarcation of the Rotterdam case study area.

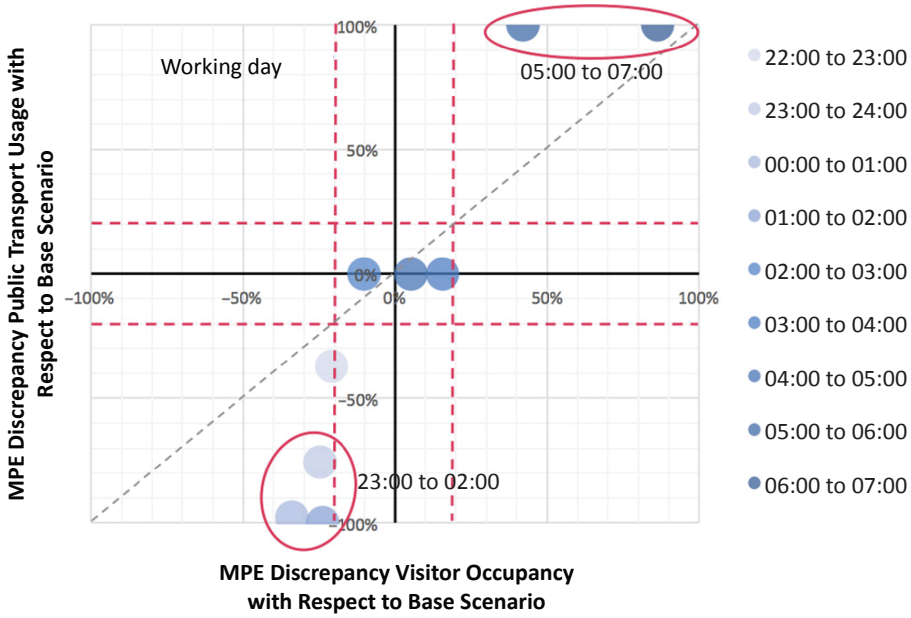

FIGURE 4 Discrepancies of public transport usage and visitor occupancy (hourly; working day) compared with the base scenario.

decreased on an hourly basis, with the latter decreasing much more sharply than the former. This may suggest that the service ended too early, given that reductions in the level of visitors who were using travel modes other than public transport exhibited a slower pace. During the night, between 2:00 and 5:00 a.m., no significant changes were observed; in the early mornings, from 5:00 a.m. onward, a rapid increase in both public transport usage and visitor occupancy was measured. At the aggregate level, the transition from the night to the daytime network seems to be justified, given the simultaneous change in overall occupancy levels.

Zonal Hourly Analysis Results

Investigating hourly changes at the zone level allowed identification of zones where the night service might be inadequate. The results indicated that visitor occupancy continued to change in several zones in the late evening after public transport services ceased (1:00 to 2:00 a.m.) and it increased considerably in the early morning, when the services are gradually resumed (5:00 to 6:00 a.m.). 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 were available for the working days. The minimum MPE value for the smartcard data is −100%; 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 over no operations in the hour before. The maximum MPE value for GSM is +88%. Light red and light green imply a decrease or increase within the threshold value range—that is, the range defined as a nonsignificant change. Zones with contradictory colors were of particular interest. For the time interval 1:00 to 2:00 a.m. (Figure 5*a*), zones of interest were 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 because transport operations have already stopped. For the time interval 5:00 to 6:00 a.m. (Figure 5*b*), zones of interest are found especially in the western and southern parts of the case study area (dark green background, light red dot). These industrial and logistic zones around the large port area already had a significant increase in visitor occupancy during this hour relative to the previous hour, although public transport operations had not resumed yet.

The MPE calculates relative changes to allow relating the two data sources to each other. The net absolute change in visitor occupancy compared with the previous hour is also of interest to the local operator to help in assessing the magnitude of the potential demand. Tables 1 and 2 summarize the relative change of both public transport usage and visitor occupancy for time intervals 1:00 to 2:00 a.m. and 5:00 to 6:00 a.m. compared with the previous hour for the zones of interest, identified on the basis of 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 many 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, because the change in occupancy corresponds to the net change, indicating a potential for public transport services during these hours. It is especially important to cater for this demand because of the alcohol consumption that is customary in nightlife.

During the early morning, between 5:00 and 6:00 a.m., a net change of 1,600 in visitor occupancy was observed in Barendrecht (Table 2 and 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, that is, Schiedam and Vlaardingen, a relatively high absolute value of visitor occupancy was observed, suggesting that there is a potential demand for additional public transport in the early mornings.

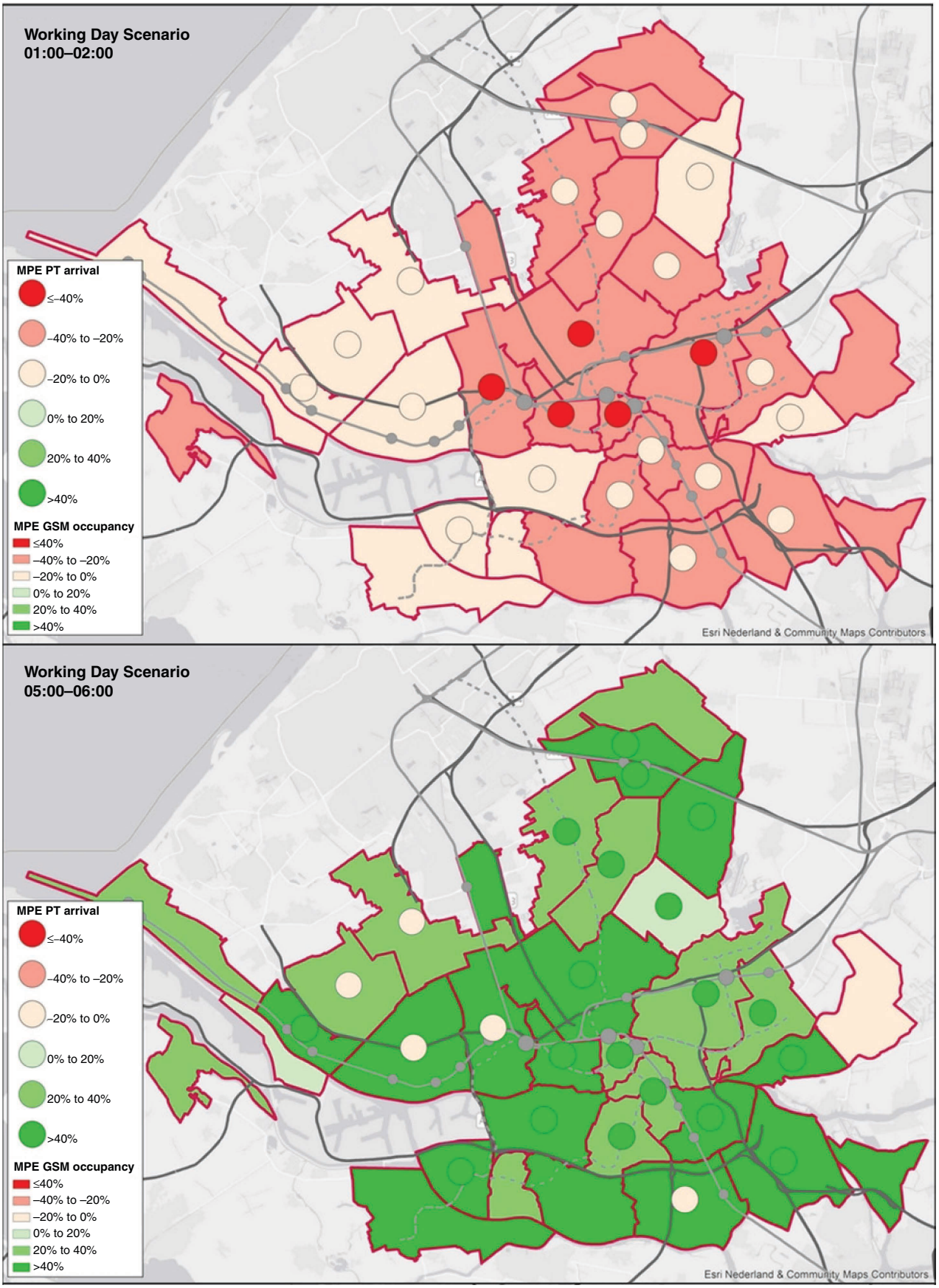

FIGURE 5 MPE of hourly values on working days for visitor occupancy (background) and public transport usage (circles) for (*top*) 1:00 to 2:00 a.m. and (*bottom*) 5:00 to 6:00 a.m.

TABLE 1 Relative and Absolute Changes in Visitor Occupancy for Selected Zones from 1:00 to 2:00 a.m. on Working Days

Zone Name	MPE of Visitor Occupancy (%)	Net Change in Visitor Occupancy
Barendrecht	–27	550
Bergschenhoek	–23	200
Berkel and 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 from 5:00 to 6:00 a.m. 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

CONCLUSIONS AND RECOMMENDATIONS

The public transport industry faces challenges in catering for the variety of mobility patterns and corresponding needs and preferences of passengers. Although data fusion can be used to investigate spatial and temporal variations in travel demand, it is only seldom used by public transport operators. This research developed a methodology to fuse smartcard and GSM data to allow analysis of public transport usage in relation to overall travel demand. On the basis 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.

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

The data fusion approach proposed here can be used to explore and fuse a large range of data sets that contain information (in aggregated or disaggregated form) for origins or destinations, or both, in transport networks. The potential transit ridership for night service, a period for which conventional onboard surveys and travel diaries offer no information because of its small share and distinctive travel patterns, was investigated. Fusing smartcard and GSM data, which are readily available and collected continuously on a large scale, allows analyzing travel demand for periods, circumstances, and areas for which conventional data collection efforts will not yield sufficient data. This includes, for example, large-scale events, service disruptions, various weather conditions, and low-demand origin–destination pairs, which were examined in a follow-up study. The approach adopted in this research can be thus used in a large range of applications where demand data availability may otherwise undermine the analysis.

Several limitations of the methodology can be identified, pertaining to data processing issues. Even though ongoing efforts are decreasing 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 relationships in the GSM data would provide information on the direction of the potential public transport demand. Smartcard data in the Netherlands are owned and stored by individual public transport operators. Fusing data from different operators, including the national railway, will enable identification of passengers transferring between services provided by different operators.

ACKNOWLEDGMENTS

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

REFERENCES

1. Guedes, M. C. M., N. Oliveira, S. Santiago, and G. Smirnov. On the Evaluation of a Public Transportation Network Quality: Criteria Validation Methodology. *Research in Transportation Economics,* Vol. 36, No. 1, 2012, pp. 39–44. https://doi.org/10.1016/j.retrec.2012.03.013.
2. Cats, O., Q. Wang, and Y. Zhao. Identification and Classification of Public Transport Activity Centers in Stockholm Using Passenger Flows Data. *Journal of Transport Geography,* Vol. 48, 2015, pp. 10–22. https://doi.org/10.1016/j.jtrangeo.2015.08.005.
3. Gutierrez, J., and J. C. Garcia-Palomares. New Spatial Patterns of Mobility Within the Metropolitan Area of Madrid: Toward More Complex and Dispersed Flow Networks. *Journal of Transport Geography,* Vol. 15, No. 1, 2007, pp. 18–30. https://doi.org/10.1016/j.jtrangeo.2006.01.002.
4. Pelletier, M. P., M. Trepanier, and C. Morency. Smart Card Data Use in Public Transit: A Literature Review. *Transportation Research Part C: Emerging Technologies,* Vol. 19, No. 4, 2011, pp. 557–568. https://doi.org/10.1016/j.trc.2010.12.003.
5. Elias, D., F. Nadler, J. Stehno, J. Krosche, and M. Lindorfer. SOMOBIL—Improving Public Transport Planning Through Mobile Phone Data Analysis. *Transportation Research Procedia,* Vol. 14, 2016, pp. 4478–4485. https://doi.org/10.1016/j.trpro.2016.05.370.

6. Durand, C. P., X. Tang, K. P. Gabriel, I. N. Sener, A. O. Oluyomi, G. Knell, A. K. Porter, D. M. Hoelscher, and H. W. Kohl III. The Association of Trip Distance with Walking to Reach Public Transit: Data from the California Household Travel Survey. *Journal of Transport and Health,* Vol. 3, No. 2, 2016, pp. 154–160. https://doi.org/10.1016/j.jth.2015.08.007.
7. Long, Y., and J. C. Thill. 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. https://doi.org/10.1016/j.compenvurbsys.2015.02.005.
8. Del Castillo, J. M., and F. G. Benitez. 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. https://doi.org/10.1016/j.sbspro.2012.09.825.
9. Frias-Martinez, V., C. Soguero, and E. Frias-Martinez. Estimation of Urban Commuting Patterns Using Cellphone Network Data. In *Proceedings of 6th Transport Research Arena,* April 18–21, 2016, Warsaw, Poland.
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, No. 33, 2015, pp. 369–389. https://doi.org/10.1145/2346496.2346499.
11. Van der Mede, P. *Over Het Meten van Mobiliteit met GSM-data: Mogelijkheden en Onmogelijkheden.* Contribution for Colloquium Vervoersplanologisch Speurwerk, Eindhoven, Netherlands, 2014 (in Dutch).
12. Aguilera, V., S. Allio, V. Benezech, F. Combes, and C. Milion. Using Cell Phone Data to Measure Quality of Service and Passenger Flows of Paris Transit System. *Transportation Research Part C: Emerging Technologies,* Vol. 43, No. 2, 2014, pp. 198–211. https://doi.org/10.1016/j.trc.2013.11.007.
13. Calabrese, F., M. Diao, G. Di Lorenzo, J. Ferreira, Jr., and C. Ratti. 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. https://doi.org/10.1016/j.trc.2012.09.009.
14. Iqbal, M. S., C. F. Choudhury, P. Wang, and M. C. Gonzalez. Development of Origin-Destination Matrices Using Mobile Phone Call Data. *Transportation Research Part C: Emerging Technologies,* Vol. 40, 2014, pp. 63–74. https://doi.org/10.1016/j.trc.2014.01.002.
15. Kusakabe, T., and Y. Asakura. Behavioral Data Mining of Transit Smart Card Data: A Data Fusion Approach. *Transportation Research Part C: Emerging Technologies,* Vol. 46, 2014, pp. 179–191. https://doi.org/10.1016/j.trc.2014.05.012.
16. Elfrink, M., M. Courtz, S. Metz, M. Ebben, and J. Weppner. *OV-Potentie Opsporen Door Datafusie.* Nationaal Verkeerskundecongres, 2015 (in Dutch).
17. Holleczek, T., L. Yu, J. K. Lee, O. Senn, C. Ratti, and P. Jaillet. 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). Association for Computing Machinery, 2014.
18. Duff-Riddell, W. R., and C. J. Bester. Network Modeling Approach to Transit Network Design. *Journal of Urban Planning and Development,* Vol. 131, No. 2, 2005, pp. 87–97. https://doi.org/10.1061/(ASCE)0733-9488(2005)131:2(87).
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*. MS thesis. Delft University of Technology, 2016.
20. ViewDAT. *ViewDAT.* http://view.dat.nl/viewdat/. Accessed October 2015.
21. Alexander, L., 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: Emerging Technologies,* Vol. 58, 2015, pp. 240–250. https://doi.org/10.1016/j.trc.2015.02.018.
22. Liu, L., A. Hou, A. Biderman, C. Ratti, and J. Chen. Understanding Individual and Collective Mobility Patterns from Smart Card Records: A Case Study in Shenzhen. In *Proceedings of the 12th International IEEE Conference on Intelligent Transportation Systems*, St. Louis, Mo., October 3–7, 2009, pp. 1–6. https://doi.org/10.1109/ITSC.2009.5309662.
23. Nishiuchi, H., J. King, and T. Todoroki. Spatial-Temporal Daily Frequent Trip Pattern of Public Transport Passengers Using Smart Card Data. *International Journal of Intelligent Transportation Systems Research,* Vol. 11, No. 1, 2013, pp. 1–10. https://doi.org/10.1007/s13177-012-0051-7.
24. RET. *Over RET,* http://corporate.ret.nl (In Dutch). Accessed July 2016.

The Public Transportation Group peer-reviewed this paper.