

Forecasting Transport Mode Use with Support Vector Machines Based Approach

Ivana Semanjski^a, Angel J. Lopez^{a,b}, Sidharta Gautama^a

Since information and communication technologies have become an integral part of our everyday lives, it only seems logical that the smart city concept should attempt to explore the role of an integrated information and communication approach to city asset management and raising the quality of life of its citizens. Raising the quality of life relies not only on improving the management of a city's systems (e.g. transportation system) but also on the provision of timely and relevant information to its citizens to allow them to make better informed decisions. This requires the use of forecasting models. In this paper, a support vector machine-based model is developed to predict future mobility behavior from crowdsourced data. Crowdsourced data are collected through a dedicated smartphone app tracking mobility behavior. The use of a forecasting model of this type can facilitate the management of a smart city's mobility system while simultaneously ensuring the timely provision of relevant pre-travel information to its citizens.

KEY WORDS

- ~ Component
- ~ Travel behavior
- ~ Smart city
- ~ Crowdsourcing
- ~ GNSS
- ~ Smartphones
- ~ Transport mode
- ~ Forecasting
- ~ Support vector machines
- ~ Pre-travel information service.

a. Ghent University, Department of Telecommunications and Information Processing, Ghent, Belgium

e-mail: ivana.semanjski@ugent.be

b. Facultad de Ingeniería en Electricidad y Computación, Politécnica del Litoral, Guayaquil, Ecuador

e-mail: alopez@espol.edu.ec

1. INTRODUCTION

The influence of information and communication technologies (ICT) on transport planning is nothing new. On the one hand, transport planning involves multiple complex models attempting to generalize the dynamic features of human and cargo movement, the need for mobility and predict the future state of the system. On the other hand, the ICT are in a continuous state of development, with ever increasing processing powers and calculation capabilities, innovative protocols and a growing number of available sensors. Thus, the integration of ICT tools into the transport planning process is doubly challenging and requires constant synchronization of available technology and transport planning needs. Much has been done in this field over the last decades and particularly in recent years, when the integration of the role of ICT at a higher level enabled the development of smart cities (Tranos and Gertner, 2012; Neirotti et al., 2014; Marsa-Maestre et al., 2008; Beswick, 2014; Alonso and Rossi, 2011; Atzori et al., 2010; Nam and Pardo, 2011; The Climate Group, 2011). In this context, location acquisition technologies are an important basis for smart city applications (Lazaroiu and Roscia, 2012; Lu and Liu, 2012). As an urban development concept, smart cities aim to safely incorporate ICT solutions into city asset management (e.g. local departments' information systems, transportation systems, hospitals, power plants, water supply networks, waste management etc.). The purpose of building a smart city is to raise the quality of life of its inhabitants by using technology to improve the efficiency of services and enable both policy makers and citizens to make better informed decisions. When it comes to the transportation aspect of smart cities, location information acquisition is often supported by Global Navigation Satellite System (GNSS) data. The GNSS is a constellation of satellites providing signals from space transmitting positioning and timing data to GNSS receivers. The receivers then use this data to pinpoint locations. Probably the best know GNSS

system is the Global Positioning System (GPS) developed by USA's NAVSTAR, but other systems like Russian Global'naya Navigatsionnaya Sputnikovaya Sistema (GLONASS) and Chinese BeiDou Navigation Satellite System are also operational, while others still, like European Union's Galileo system, are on their way to provide global coverage. Literature provides interesting examples of GNSS use in the extraction of origin-destination (OD) matrices (Pulselli et al., 2006; Novak et al., 2013; Järv et al., 2014; Iqbal et al., 2014), validation of travel behaviour models (Liu et al., 2014; Yuan et al., 2012) or rush hour analysis (Bar-Gera, 2007; Järv et al., 2012). Furthermore, when analysing the use of GNSS data in mobility studies, we can differentiate between the implementation of (a) a dedicated GNSS sensor and (b) an integrated GNSS sensor. The former usually refers to a dedicated sensor placed into a vehicle or a portable GNSS device carried by an individual to log his mobility behaviour. These studies are usually based on limited samples (e.g. due to the degree of discipline required from respondents to carry the device with them) or take only some modes of transportation into consideration (e.g. the device only tracks motorised vehicles). The latter involves the embedding of GNSS chipsets into devices that are not primary dedicated to positioning. Probably the most common device of this type is a mobile phone (Semanjski and Gautama, 2016). Since mobile phones have several integrated sensors (e.g. cameras, accelerometers etc.) and we usually carry them with us without giving it a second thought, they have the potential to overcome the above mentioned limitations and provide a variety of valuable datasets for mobility studies. These

datasets are most frequently referred to as crowdsourced data. Nevertheless, little is known about the potential of crowdsourced data for smart city mobility management and even less about the context of personalized mobility services and the interaction between a city and the users of its transportation system. Since only scarce literature tackles this idea (Semanjski and Gautama, 2015; Semanjski et al., 2016), the contribution of this paper to this line of research is the use of crowdsourced data from smartphones and a support vector machine algorithm to forecast the mode of transportation one will use for an upcoming trip. Forecasting is based on a set of given conditions (location, purpose of the trip, time of day, etc.). This could potentially serve as the basis for an application which could be used to indicate relevant intended mobility behaviour in a timely fashion, allowing the obtainment of relevant transportation information and supporting the process of making better informed mobility related decisions. Furthermore, these information could also be used to support smart city mobility management by ensuring timely management of transportation-related activities and services.

2. METHOD AND DATA

2.1. Data Collection

Data on mobility behavior was collected using an Android smartphone application, Routecoach (Ghent University, 2016), developed at Ghent University in Belgium. The Routecoach

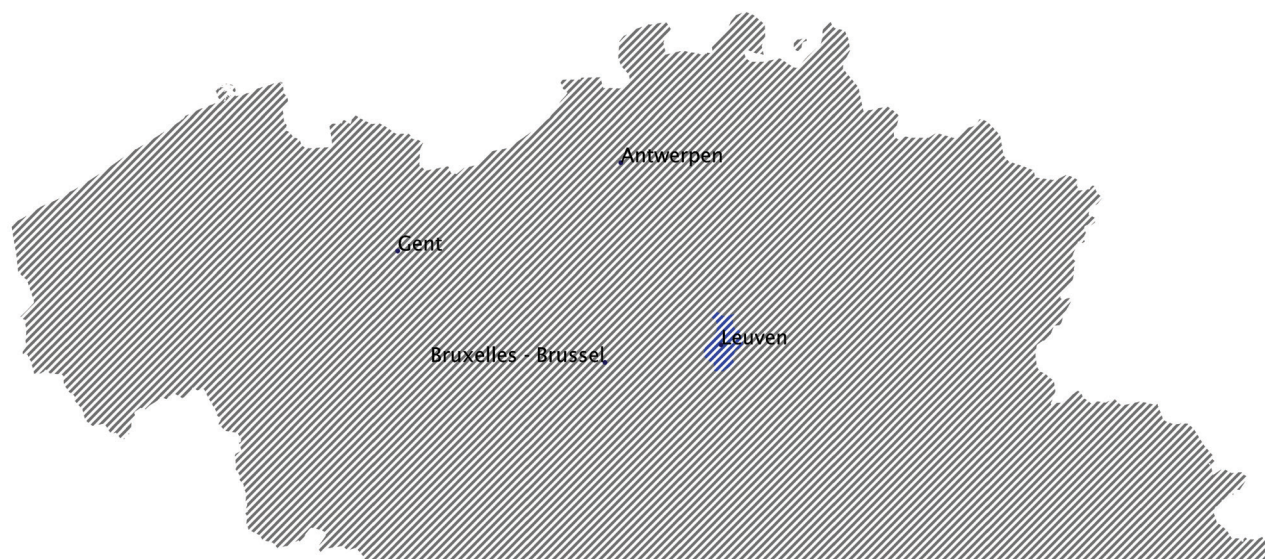


Figure 1.
Location of the city of Leuven in the Flanders region.

application was a part of the sustainable mobility campaign in the Flemish-Brabant province. The main goal of the campaign was to develop a transferable evaluation and planning toolkit for mobility projects capable of being adopted by planners (New Integrated Smart Transport Options, 2016). The data collection

process lasted from January to April 2015. A total of 8,303 users actively participated by downloading the freely available application and collecting the data on more than 30,000 trips.

In our analysis, we used a part of the overall dataset that included 'interactively' (also called 'actively') (Semanjski and

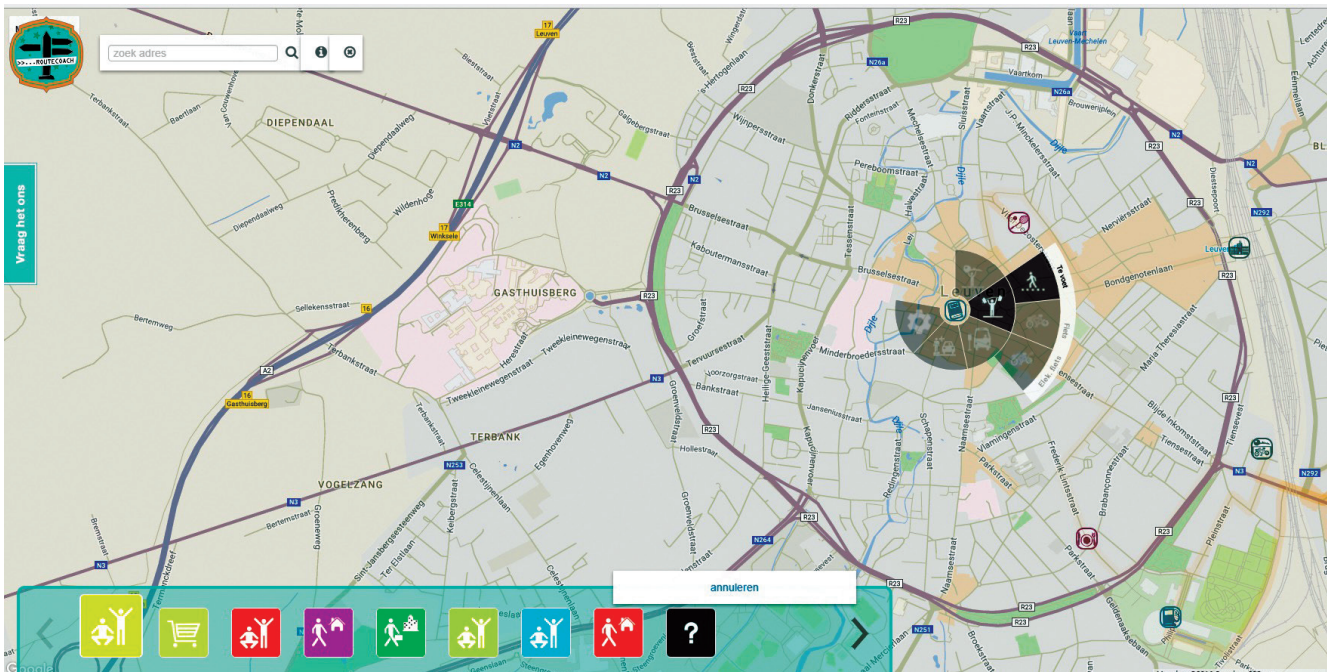


Figure 2. Geo interface for data validation and quality control.

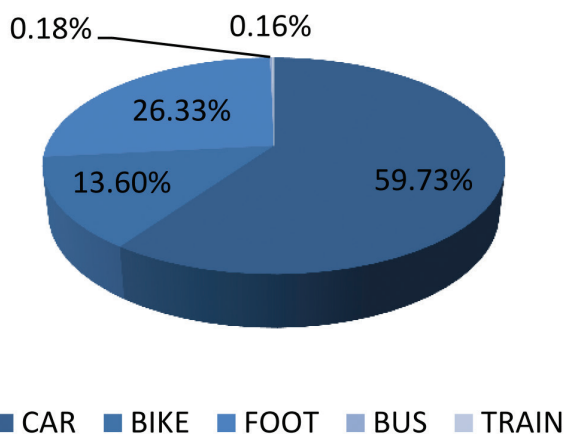


Figure 3. Modal split (validated trips).

Gautama, 2016) logged trips in the city of Leuven. The city of Leuven (Figure 1) is the capital of the Flemish-Brabant province, located about 25 kilometers east of Brussels (the capital of Belgium). The two cities are well-connected by roads, railway and a bike highway. Leuven is a very dynamic city with a lot of traffic and related traffic congestion. The use of public transportation (only bus is available) in the city has seen a fivefold increase in the last 20 years. Also, cycling is a very popular mode of transportation, with the share of 17-20% (New Integrated Smart Transport Options, 2016).

Our sub-dataset consists of 17,040 validated trips created by 292 individuals, meaning that each individual made around 60 trips in average. The greatest number of trips were made by car and the smallest by public transportation (Figure 3). The distribution of validated trips over 24 hours clearly indicates morning and afternoon peaks (Figure 4).

Detailed description of attributes collected for each trip is given in Table 1.

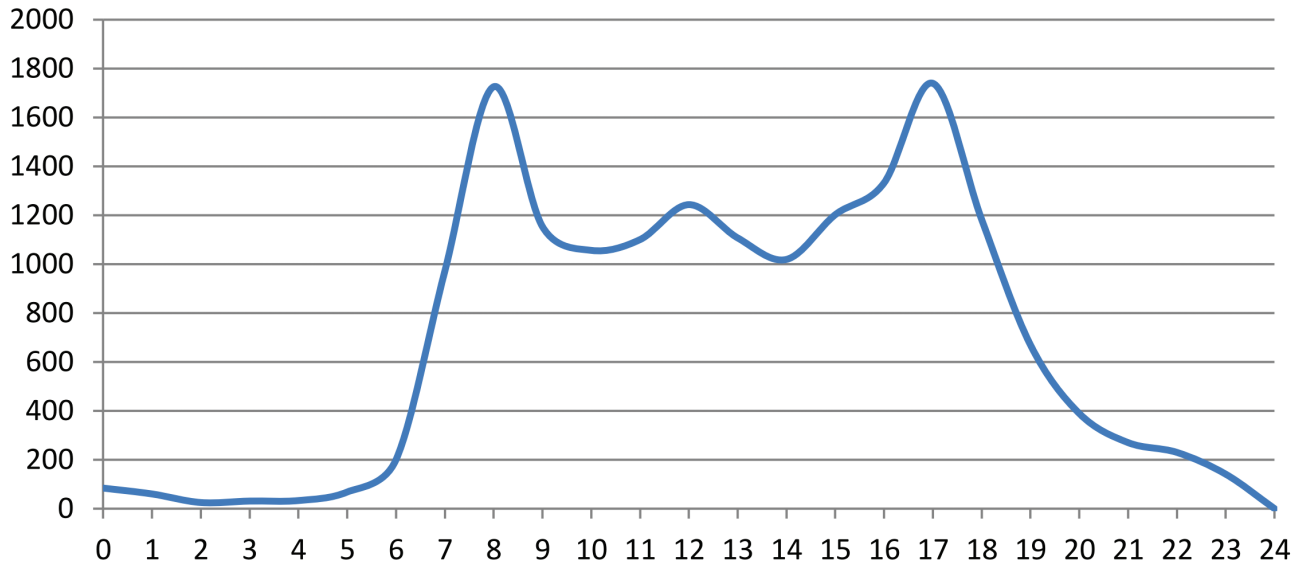


Figure 4.
Modal split (validated trips).

Table 1.
Description of variables.

Variable	Acronym	Description
User's ID	userid	Unique identifier of the user/device
Trip's ID	tripid	Unique identifier of the trip
Trip's start time	starttime	Year, month, day, hour, minute and second when the trip started
Trip's stop time	stoptime	Year, month, day, hour, minute and second when the trip ended
Trip's start location	startpoint	Geographic location of the trip's starting point
Trip's end location	endpoint	Geographic location of the trip's destination point
Distance	distance	Distance between the trip's starting and destination points measured in kilometres
Mode of transportation	transportmode	Mode of transportation used for the trip
Purpose of the trip	purpose	The purpose of the trip made (go to work, shopping, recreation, school...)
Working day identification	weekday	Boolean value indicating whether the trip started on a working day
Holiday identification	weekend	Boolean value indicating whether the trip started on a holiday or weekend

2.2. Classification of Support Vector Machines

Support vector machines (SVM) are a supervised machine learning algorithm based on the concept of decision hyperplanes. The decision hyperplanes separate observations

that have different class membership. The observations can frequently not be easily separated, and mathematical functions (kernels) are used to map them in order to differentiate between the classes. First, the training dataset is used. The training dataset is a dataset in which class membership is known in advance for

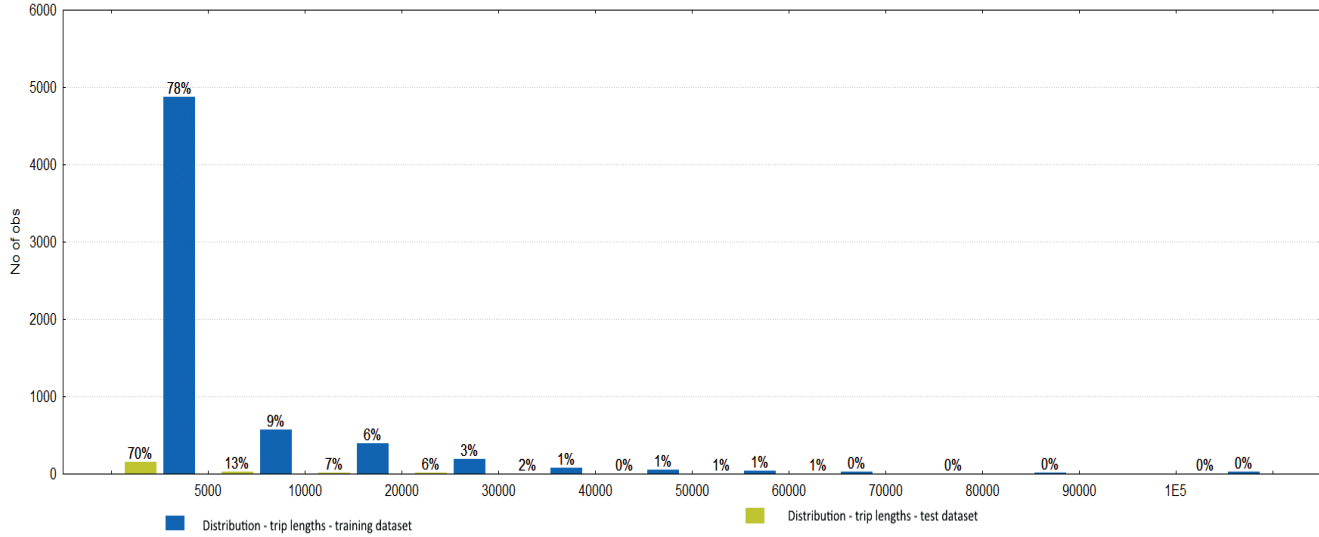


Figure 5. Success rate in relation to the trip lengths.

each observation. The purpose of this dataset is to train the model and select the best possible decision hyperplanes. Then a new dataset is introduced (test dataset) to test the success rate of the first step. This dataset also has known class membership and serves for objective evaluation of the model (since model was not trained with it). A more detailed overview of SVM algorithm can be found in literature (Čavar et al., 2011; Joo et al., 2015; Vlahogianni, 2015).

In our study, we divided the complete dataset into two parts; 75 % of the dataset was used as training and 25 % as test dataset. Sampling was random and an insight into the distribution of trip lengths between training and test dataset reveals a quite balanced representation of different trip lengths in both samples (Figure 5). The input dataset consists of trip observations for each individual, with each trip considered to be a distance between two locations passed by a single mode of transportation. Each trip is described with variables listed in Table 1.

For the SVM classification we applied the C-SVM type. The forecasting minimization error function for the applied C-SVM is defined as:

$$\frac{1}{2} w^T w + C \sum_{i=1}^N \xi_i \quad (1)$$

subject to constraints:

$$Y_i (w^T \Phi(x_i) + b) \geq 1 - \xi_i \quad (2)$$

$$\xi_i \geq 0 \quad (3)$$

Where $i=1, \dots, N$, C is the capacity constant, w represents the vector of coefficients, b is a constant, and ξ_i are parameters for handling non-separable inputs and Φ stands for kernel function. Kernel function used in our example is radial basis function that transforms input to the feature space as defined by equation (4):

$$K(X_i, X_j) = \Phi(X_i) \cdot \Phi(X_j) = X_i X_j \quad (4)$$

The value of parameters C and γ is defined in the training phase based on the results of tenfold cross-validation. The obtained values were 3 for C and 0.2 for γ . The output variable of the model was the mode of transportation to be used for the next trip and forecasting time frame was one hour.

3. RESULTS

The overall success rate of the forecasting model was 82 % (Table 2). The most challenging part was to create decision hyperplanes capable of differentiating between trips made by personal car, bike and public transportation (bus), since this resulted in more than 1,000 support vectors for bike and car transport modes.

Considering each mode of transportation separately (Figure 6), the use of a personal vehicle was the easiest to predict. A detailed look into the confusion matrix reveals the occurrence of miss-classifications between the bike, car and foot modes of transportation (Figure 7). In most cases, trips predicted to be made by car were forecasted to be made by bike or on foot. Potentially, taking weather conditions into account might provide a wider context for the miss-classifications and the phase model could be extended to integrate such insights. Furthermore, Figure 8 and 9 provide more details on the purpose

Table 2.
Model results.

Kernel type	Radial Basis Function
Classification accuracy	81.87 %
Number of SVs	2921 (1 bounded)
Number of SVs (BIKE)	1187
Number of SVs (BUS)	687
Number of SVs (CAR)	1033
Number of SVs (FOOT)	5
Number of SVs (TRAIN)	9

of the forecasted trips. Quite different distributions indicate the importance of the availability of information on the purpose of the trip for predicting the mode of transportation to be used.

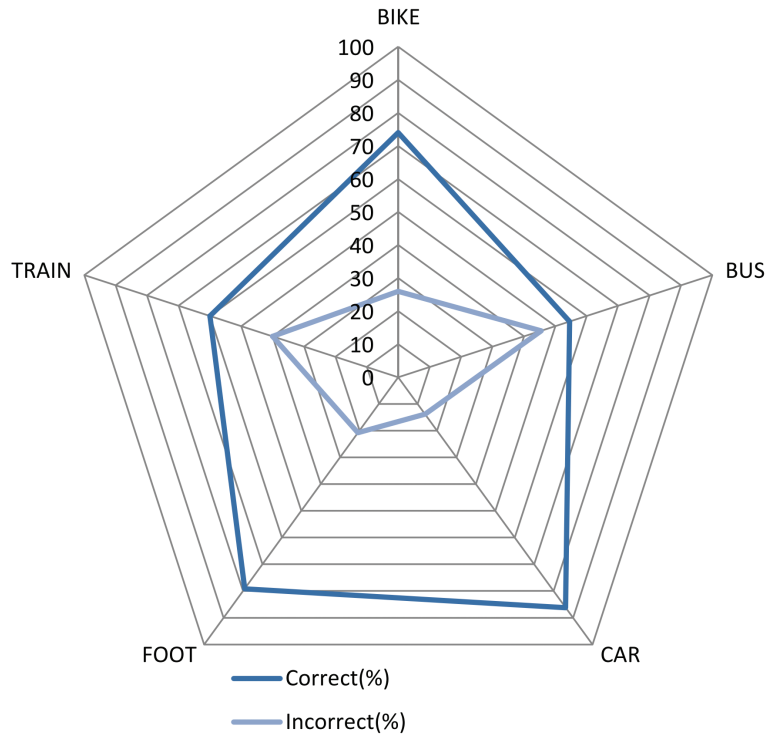


Figure 6.
Model's success rate by modes of transportation.

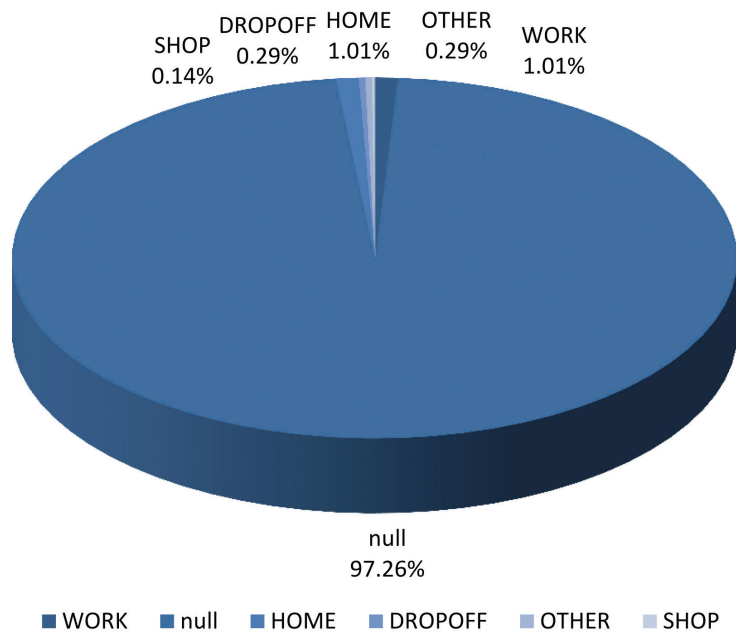


Figure 7. Purpose of trips for which mode of transportation was incorrectly forecasted.

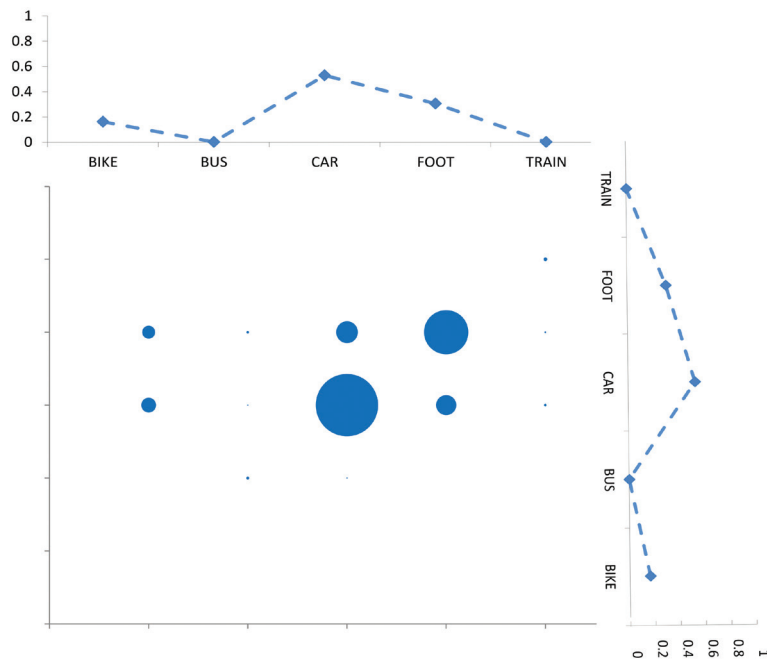


Figure 8. Confusion matrix.

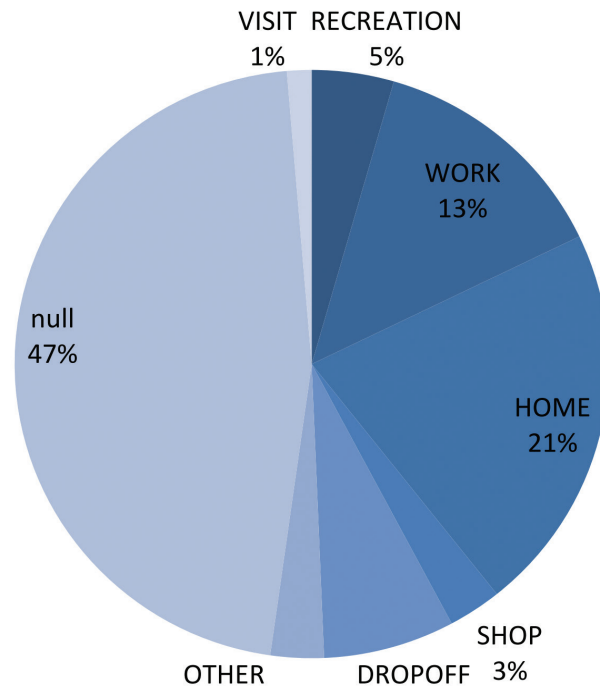


Figure 9. Purpose of trips for which mode of transportation was correctly forecasted.

4. DISCUSSION

Although the implemented SVM model has quite a high success rate (82 %), there is still room for the improvement of forecasting results by additional extensions and considerations based on the gained insights. Firstly, since the model yielded a correct forecast in just above eight out of ten trips, when implemented to provide pre-travel information, it could be useful to provide information on two modes of transportation most likely to be used. This could increase the relevance of model outcomes to the user and have a decisive role when the user is considering more than one option. Furthermore, this would also allow the integration of such information into the city's transport management system and their coordination with the city's preferences. For example, if the city wishes to promote a new public transportation line or bike route, it could give these options additional weight in the model so that this information is provided to the user whenever a feasible route using promoted options exists. In addition, the highest degree of confusion was created between the use of a personal vehicle and active modes of transportation like bike and foot, which are at the far ends of the sustainable mobility spectrum. The context in which miss-

classified trips occurred is worth further study in future research. Potential reasons include unfavorable weather conditions when users are more prone to use personal vehicles and different trip purposes (e.g. one is more likely to use a personal vehicle to drop-off multiple family members than for recreation purposes). The former reason can be successfully examined by fusing weather data with the crowdsourced dataset. It could provide a broader context for the occurrence of confusion. The latter requires knowledge of the purpose of the trip. As forecasting timeframe was one hour, the purpose of the trip can be presumed to be known in advance. However, unless this information is indicated by the user, the purpose of the trip must also be forecasted, which contributes to the complexity of the model and might influence the success rate of the forecast. Therefore, it could be beneficial to examine the complementarity of trip purpose forecasting models with transport mode forecasting in more detail and evaluate added value they give to each other.

Compared with results from literature (Semanjski and Gautama, 2015) where gradient boosting trees were used to forecast mode of transportation to be used for the next trip, our results achieved with the support vector machines-based model have a 10 % higher success rate. This shows the potential for

the expansion of the support vector machines-based model to incorporate other data sources and be successfully implemented to support smart city mobility planning and managing process.

5. CONCLUSION

The support vector machines-based model had the success rate of 82 % in forecasting the mode of transportation to be used for the next trip. The implementation of this model into smart city mobility system management and planning processes clearly has high potential, since it can result in the development of a more advanced pre-travel information service. Furthermore, the insights gained are already indicative of potential future expansions of the model which could increase the usability of the output results and their relevance to end users.

ACKNOWLEDGMENT

This research is funded by INTERREG North-West Europe project New Integrated Smart Transport Options (NISTO), the Flemish government agency for Innovation by Science and Technology and the Flemish Institute for Mobility.

REFERENCES

- Alonso, G. J. A. and Rossi, A., (2011), New Trends for Smart Cities, Bezons: ATOS.
- Atzori, L., Iera, A. and Morabito, G., (2010), The Internet of Things: A Survey, *Computer Networks*, 54(15), pp. 2787–2805., <http://dx.doi.org/10.1016/j.comnet.2010.05.010>
- Bar-Gera, H., (2007), Evaluation of a Cellular Phone-based System for Measurements of Traffic Speeds and Travel Times: A Case Study from Israel, *Transportation Research Part C: Emerging Technologies*, 15(6), pp. 380–391., <http://dx.doi.org/10.1016/j.trc.2007.06.003>
- Beswick, S., (2014), *Smart Cities in Europe: Enabling Innovation*, London: Osborne Clarke.
- Ćavar, I., Kavran, Z. and Petrović, M., (2011), Hybrid Approach for Urban Roads Classification Based on GPS Tracks and Road Subsegments Data, *Promet*, 23(4), pp. 289 – 296., <http://dx.doi.org/10.7307/ptt.v23i4.131>
- Ghent University, "Routecoach," Google Play, available at: <https://play.google.com/store/apps/details?id=com.move.routecoach>, [accessed 27 May 2016].
- Iqbal, S., Choudhury, C. F., Wang, P. and González, M. C., (2014), Development of Origin-destination Matrices Using Mobile Phone Call Data, *Transportation Research Part C: Emerging Technologies*, 40, pp. 63–74., <http://dx.doi.org/10.1016/j.trc.2014.01.002>
- Järv, O., Ahas, R., Saluveer, E., Derudder, B. and Witlox, F., (2012), Mobile Phones in a Traffic Flow: A Geographical Perspective to Evening Rush Hour Traffic Analysis Using Call Detail Records, *PLoS ONE*, 7(11), e49171., <http://dx.doi.org/10.1371/journal.pone.0049171>
- Järv, O., Ahas, R. and Witlox, F., (2014), Understanding Monthly Variability in Human Activity Spaces: A Twelve-month Study Using Mobile Phone Call Detail Records, *Transportation Research Part C: Emerging Technologies*, 38, pp. 122–135., <http://dx.doi.org/10.1016/j.trc.2013.11.003>
- Joo, S., Oh, C., Jeong, E. and Lee, G., (2015), Categorizing Bicycling Environments Using GPS-based Public Bicycle Speed Data, *Transportation Research Part C: Emerging Technologies*, 56, pp. 239 – 250., <http://dx.doi.org/10.1016/j.trc.2015.04.012>
- Lazaroiu, G. C. and Roscia, M., (2012), Definition Methodology for the Smart Cities Model, *Energy*, 47(1), pp. 326 – 332., <http://dx.doi.org/10.1016/j.energy.2012.09.028>
- Liu, F., Janssens, D., Cui, J., Wang, Y., Wets, G. and Cools, M., (2014), Building a Validation Measure for Activity-based Transportation Models Based on Mobile Phone Data, *Expert Systems with Applications*, 41(14), pp. 6174 –6189., <http://dx.doi.org/10.1016/j.eswa.2014.03.054>
- Lu, Y. and Liu, Y., (2012), Pervasive Location Acquisition Technologies: Opportunities and Challenges for Geospatial Studies, *Computers, Environment and Urban Systems*, 36(2), pp. 105–108., <http://dx.doi.org/10.1016/j.compenvurbsys.2012.02.002>
- Marsa-Maestre, I., Lopez-Carmona, M. A., Velasco, M. A. J. R. and Navarro, I. A., (2008), Mobile Agents for Service Personalization in Smart Environments, *Journal of Networks*, 3(5), pp. 30–41.
- Nam, T. and Pardo, T. A., (2011), Conceptualizing Smart City with Dimensions of Technology, People, and Institutions, *Proc. 12th Annual International Conference on Digital Government Research*, College Park, MD, USA, June 12–15, pp. 282 – 291.
- Neirotti, P., de Marco, A., Cag, A. C., Mangano, G. and Scorrano, F., (2014), Current Trends in Smart City Initiatives: Some Stylised Facts, *Cities*, 38, pp. 25–36., <http://dx.doi.org/10.1016/j.cities.2013.12.010>
- New Integrated Smart Transport Options, "NISTO," available at: <http://www.nisto-project.eu>, [accessed 3 March 2016].
- Novak, J., Ahas, R., Aasa, A. and Silm, S., (2013), Application of Mobile Phone Location Data in Mapping of Commuting Patterns and Functional Regionalization: A Pilot Study of Estonia, *Journal of Maps*, 9(1), pp. 10–15., <http://dx.doi.org/10.1080/17445647.2012.762331>
- Official mobility statistics for Flanders, OVG, available at: <http://www.mobielvlaanderen.be/ovg/>, [accessed 27 May 2016].
- Pulselli, R. M., Ratti, C. and Tiezzi, E., (2006), City out of Chaos: Social Patterns and Organization in Urban Systems, *International Journal of Design & Nature and Ecodynamics*, 1, pp. 125–134.
- Semanjski, I. and Gautama, S., (2015), Smart City Mobility Application - Gradient Boosting Trees for Mobility Prediction and Analysis Based on Crowdsourced Data, *Sensors*, 15(7), pp. 15974–15987., <http://dx.doi.org/10.3390/s150715974>
- Semanjski, I. and Gautama, S., (2016), Sensing Human Activity for Smart Cities' Mobility Management, *InTech (In Press)*
- Semanjski, I., Lopez Aguirre, A., De Mol, J. and Gautama, S., (2016), Policy 2.0 Platform for Mobile Sensing and Incentivized Targeted Shifts in Mobility Behavior, *Sensors*, 16(7), pp. 1035., <http://dx.doi.org/10.3390/s16071035>
- The Climate Group, (2011), *ARUP, Accenture & The University of Nottingham. Information Marketplaces, The New Economics of Cities*, London, UK: The Climate Group

Tranos, E. and Gertner, D., (2012), Smart Networked Cities? Innovation: The European Journal of Social Science Research, 25(2), pp. 175–190.,
<http://dx.doi.org/10.1080/13511610.2012.660327>

Vlahogianni, E. I., (2015), Optimization of Traffic Forecasting: Intelligent Surrogate Modeling, Transportation Research Part C: Emerging Technologies, 55, pp. 14 – 23.,
<http://dx.doi.org/10.1016/j.trc.2015.03.016>

Yuan, Y., Raubal, M. , and Liu, Y., (2012), Correlating Mobile Phone Usage and Travel Behavior - A Case Study of Harbin, China, Computers, Environment and Urban Systems, 36(2), pp. 118–130.,
<http://dx.doi.org/10.1016/j.compenvurbsys.2011.07.003>