Forecasting Travel Behaviour from Crowdsourced Data with Machine Learning Based Model

Angel J. Lopez[1,2]; Ivana Semanjski[1]; Sidharta Gautama[1]
[1] Ghent University, Department of Telecomunications and Information Processing,
Ghent, Belgium
Emails: {angel.lopez, ivana.semanjski, dominique.gilles, sidharta.gautama} @ugent.be
[2]Facultad de Ingenierıa en Electricidad y Computacion
Escuela Superior Polit´ecnica del Litoral, ESPOL
Guayaquil, Ecuador
Emails: alopez @espol.edu.ec

***Abstract*—Information and communication technologies have become integral part of our everyday lives. It seems as logical consequence that smart city concept is trying to explore the role of integrated information and communication approach in managing city's assets and in providing better quality of life to its citizens. Provision of better quality of life relies on improved management of city's systems (e.g., transport system) but also on provision of timely and relevant information to its citizens in order to support them in making more informed decisions. To ensure this, use of forecasting models is needed. In this paper, we develop support vector machine based model with aim to predict future mobility behavior from crowdsourced data. The crowdsourced data are collected based on dedicated smartphone app that tracks mobility behavior. Use of such forecasting model can facilitate management of smart city's mobility system but also ensures timely provision of relevant pre-travel information to its citizens.**

Keywords-travel behavior; smart city; crowdsourceing; transport planning.

I. INTRODUCTION

Influence of information and communication technologies (ICT) on transport planning is not new. On one hand, transport planning involves multiple complex models that try to generalize dynamic features of human and cargo movements, needs for mobility and forecasts future states of the system. On the other end, ICT constantly develops, which includes higher processing powers and calculation capabilities, innovative protocols and growing number of available sensors. Thus, integration of ICT tools into transport planning process is twofold challenging and requires constant synchronization between available technology and transport planning needs. A lot has been done in this field over last decades, particularly in recent years when role of ICT is getting integrated at the higher level enabling development of smart cities [1]-[8]. In this context, location acquisition technologies play an important basis for smart city applications [9][10]. Smart cities as urban development concept aim to integrate ICT solutions in a secure fashion to manage a city's assets (e.g., local departments' information systems, transport systems, hospitals, power plants, water supply networks, waste management, etc.). The goal of building a smart city is to improve quality of life for its residents by using technology to improve the efficiency of services and to enable more informed decision making process on both policy makers' and citizens' ends. When it comes to the transport aspect of smart cities, location information acquisition is often supported by Global Navigation Satellite Systems (GNSS) data. GNSS comprehends a constellation of satellites providing signals from space that transmit positioning and timing data to GNSS receivers. The receivers then use this data to determine location. Probably the best know GNSS system is Global Positioning System (GPS) developed by USA's NAVSTAR, but other systems like Russia's Global'naya Navigatsionnaya Sputnikovaya Sistema (GLONASS) and China's BeiDou Navigation Satellite System are operational while some others are on its way to ensure global coverage like European Union's Galileo system. In the literature, there are some interesting examples of GNSS use for extraction of origin-destination (OD) matrices [11]-[14], validation of travel behaviour models [15][16] or rush hour analysis [17][18]. Furthermore, when analysing the use of GNSS data for mobility studies we can distinguish between implementation of (a) dedicated GNSS sensor and (b) integrated GNSS sensors. The first one usually includes dedicated sensor placed into a vehicle or portable GNSS device that individual carries in order to log his mobility behaviour. In these studies samples are usually limited in size (e.g., due to discipline that is required from responded in order to carry the device with him) or bias in transport modes coverage (e.g., device tracks only motorised transport modes). The second one involves integration of GNSS chipsets into devices that are not primary dedicated to location purposes. Probably the most common device of such kind today is a mobile phone [19]. As mobile phones have more sensors integrated (e.g., cameras, accelerometers, etc.) and individuals usually carry them without considering it to be a burden, they exhibit potential to overcome above

 ISBN: 978-1-61208-510-4

mentioned limitations and deliver a rich datasets for mobility studies. Most often these datasets are referred 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 interactions between a city and its transport system users. While only scarce literature tackles this idea [20][21], in this article we contribute to this line of research by using crowdsourced data from smartphones and a support vector machine algorithm to forecast transport mode one will use for the upcoming travel. Forecasting is based on a set of given conditions (location, trip's purpose, time of day, etc.). We see this as a potential application that can be used to enable timely relevant indication of intended mobility behaviour. This way relevant transport information and incentives can be provided in order to support making of more informed mobility related decisions. Furthermore smart city mobility management can be supported with this information to ensure timely management of transport related activities and services.

This paper is structured as follows: after the Introduction, Section 2 provides detailed description of the data collection procedure and adopted support vector machines approach. Section 3 summarizes the result of the transport mode forecasting procedure and is followed with discussion section. Finally, Section 5 highlights major conclusions drawn from the observed insights.

II. METHOD AND DATA

A. Data collection

Data on mobility behavior is collected via an Android smartphone application Routecoach [22], which is developed at Ghent University in Belgium. The Routecoach application was a part of sustainable mobility campaign in province of Flemish-Brabant. The main aim of the campaign was to develop an evaluation and planning toolkit for mobility projects, which is transferable and can be adopted by planners [23]. The data collection process lasted from January to April 2015. Over this period, in total, 8303 users actively participated by downloading the freely available application and collecting the data on more than 30,000 trips.

For our analysis, we used a part of overall dataset that included only 'interactively' (also called 'actively') [19] logged trips that were collected in area of city of Leuven. The city of Leuven (Fig.1) is the capital of the province of Flemish-Brabant and located about 25 kilometers east of Brussels (capital of Belgium). The two cities are well connected with road, rail and bike highway. Leuven itself is a very dynamic city that is a home to one of the largest universities in the region. The city's daily dynamics results in a lot of traffic and also related traffic congestion. This is particularly noticeable across the ring road that surrounds the city center and main road axes that bring regional traffic to the city. Potentially for this reason, the use of public transportation (only bus is available) in the city has seen a fivefold increase in last 20 years. Also, the use of active transport modes is quite common as just cycling contributes to the city's modal share with 17-20% [24].

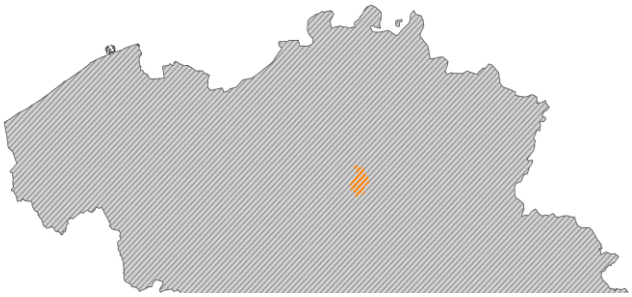

Figure 1. Location of city of Leuven in Flanders region.

As already mentioned, for our research we used only 'actively' logged trips. 'Active' logging comprehends that the logged data are validated by users. This way, through the application itself, users can confirm transport mode that they are using or indicate a reason for the trip (trip purpose) while data are being logged. In addition, for the completed trips users can perform data quality control. Data quality control is made possible over dedicated web portal (Fig. 2). Through the web portal, participants can access to their personal trip logs and use friendly, geographic information system (GIS), interface. This interface visualizes exiting trips but can also be used to report on additional trip data, correct wrongly introduced information or report on personal points of interest (like home or work locations).

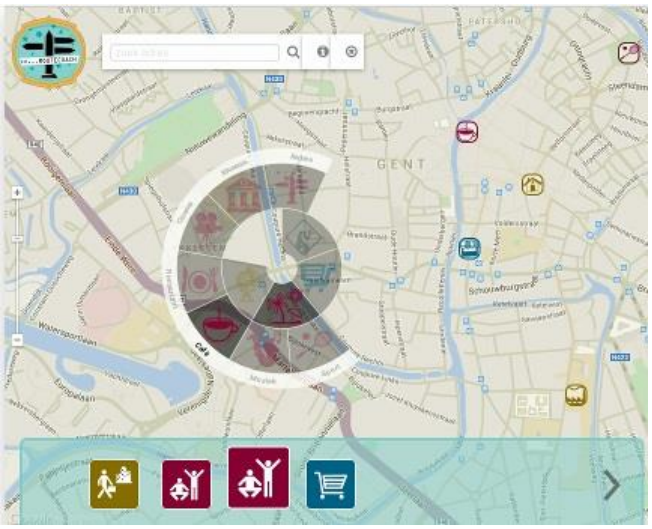

Figure 2. Geo interface for data validation and quality control.

Our sub dataset consist of 17,040 validated trips created by 292 individuals, meaning that each individual in average made around 60 trips. Most of the trips were made by car and least by public transport (Fig.3). The distribution of validated trips over 24 hours clearly indicates morning and afternoon peaks (Fig. 4).

 ISBN: 978-1-61208-510-4

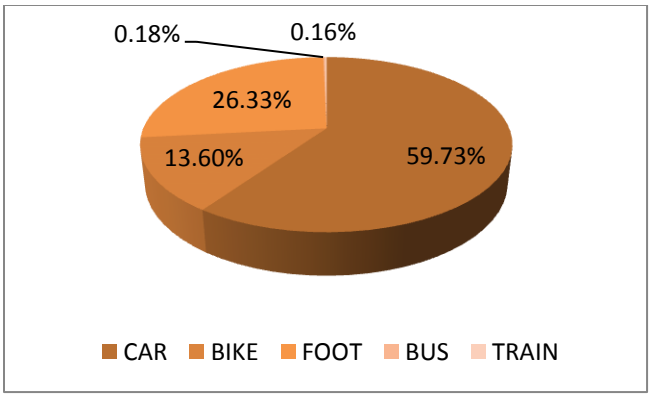

Figure 3. Modal split (validated trips).

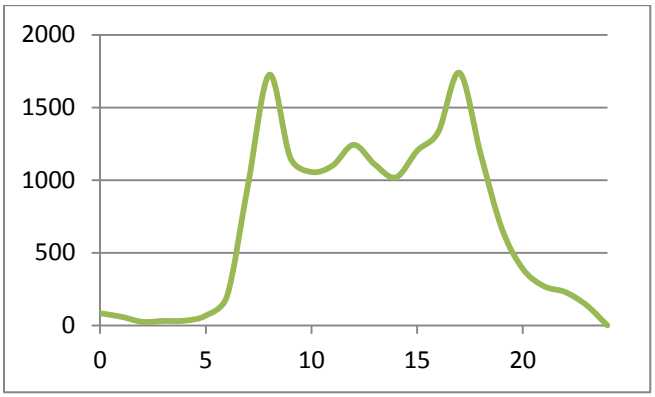

Figure 4. The distribution of validated trips over 24 hours.

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

B. Support vector machines classification

Support vector machines (SVM) are supervised machine learning algorithm that is based on the concept of decision hyperplanes. These decision hyperplanes separate between observations that have different class membership in multidimensional feature space. Quite often, it is not so easy to separate between observations, and different mathematical functions (kernels) are used to map them in order to separate among different classes. For this reason, firstly, a training dataset is introduced. The training dataset is a dataset where class membership is know in advance for each observation. This dataset serves to train the model so that, as good as possible, decision hyperplanes are selected. Later, to test the success rate of the first step, and the selected decision hyperplanes, a new dataset is introduced. This second dataset, called test dataset, also has known class membership for each observation. This knowledge of the class membership is used to objectively test the outcome of the trained model. The objectivity is based on the fact that newly introduced dataset contains observations on which the initial model was not trained, thus it allows more fair evaluation of the model results. After the selection of decision hyperplanes is confirmed on the test dataset it is considered that the model will give fair results when data with unknown class membership are introduced to it. More detailed overview of SVM algorithm can be found in literature [25]-[27].

TABLE I. 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 trip started
Trip's stop time	stoptime	Year, month, day, hour, minute and second when trip ended
Trip's start location	startpoint	Geographic location of the trip's origin point
Trip's end location	endpoint	Geographic location of the trip's destination point
Distance	distance	Distance between trip's origin and destination points measured in kilometres
Transportation mode	transportmode	Transportation mode used for the trip
Trip's purpose	purpose	The purpose of the trip made (go to work, shopping, recreation, school…)
Working day identification	week day	Boolean value that indicates if the day when trip started is a working day
Holiday identification	weekend	Boolean value that indicates if the day when trip started is a holiday or weekend

In our study, we divided complete data set in two parts; 75% has been used as training and 25% as test dataset. Sampling was random and insight into distribution of trip lengths between training and test dataset reveals quite balanced representation of different trip lengths in both samples (Fig.5). The input dataset consists of trip observations for every individual, where each trip is considered to be a path between two locations made by one transport mode. Each trip is described with variables listed in Table 1.

 ISBN: 978-1-61208-510-4

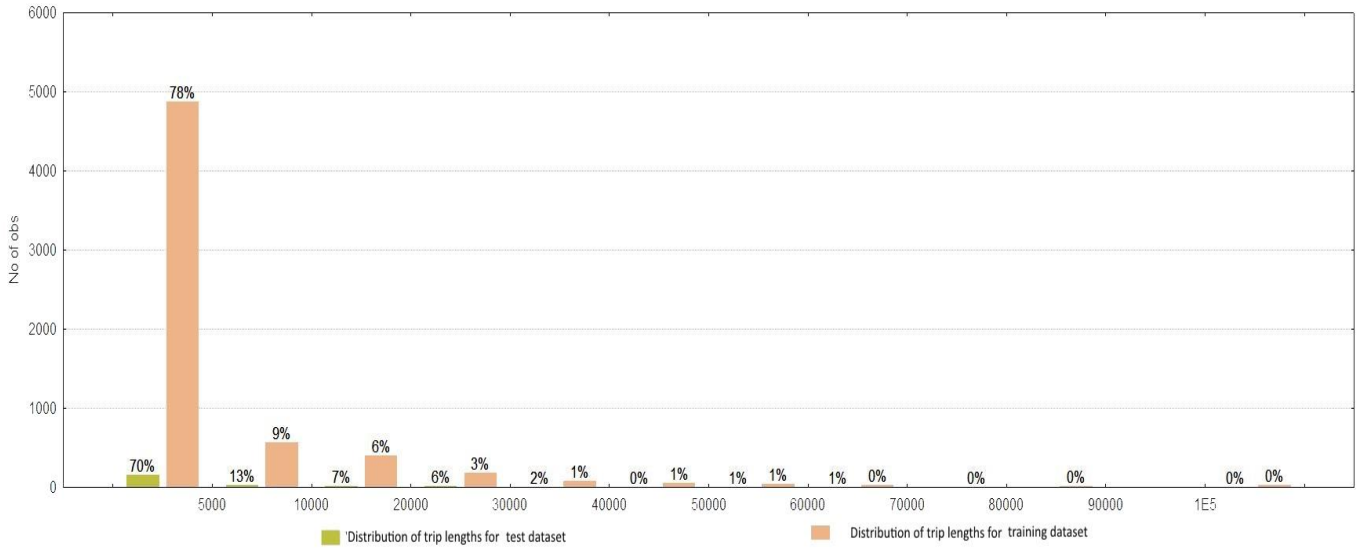

Figure 5. Success rate in relation to the trip lengths.

For the SVM classification, we applied 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 $i=1, \ldots, N$, C is the capacity constant, w represents the vector of coefficients, b is a constant, and ξi 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 (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 training phase based on the result of 10- fold cross validation. The obtained values were 3 for C and 0.2 for γ. The output variable of the model was a transport mode one will use for a next trip and forecasting time frame was one hour.

III. RESULTS

Overall success rate of the forecasting model was 82% (Table 2). The most challenging part was to create decision hyperplanes that separate between trips made with private car, bike and public transportation (bus) as this resulted in more than 1000 support vectors for bike and car transport modes.

TABLE II. 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

Considering each transport mode individually (Fig. 6), it was easiest to predict when one will use personal vehicle for the next trip. A detailed look into confusion matrix reveals occurrence of miss-classifications between transport modes bike, car and foot (Fig. 7). In most of the cases trips that were predicted to be made by car were forecasted to be made by bike or foot. Potentially, insight into weather conditions could give more details on context of miss-classifications and in future phase model can be extended to integrate these insights. Furthermore, Fig. 8 and 9 show more details on trip purposes for the forecasted trips. Quite different distributions indicate how important the availability of information on the purpose of travel is when predicting transport mode to be used for travel.

 ISBN: 978-1-61208-510-4

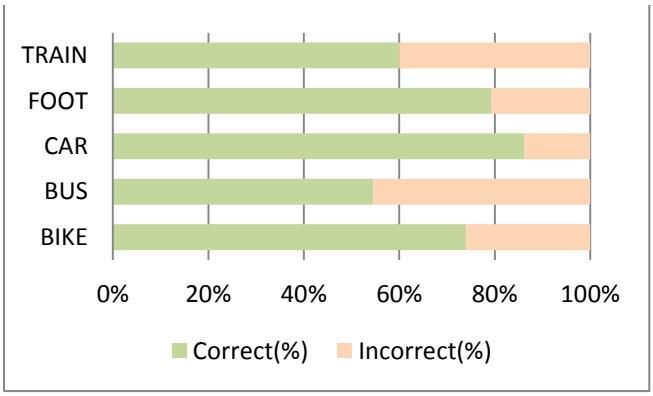

Figure 6. Model's success rate at transport mode levels

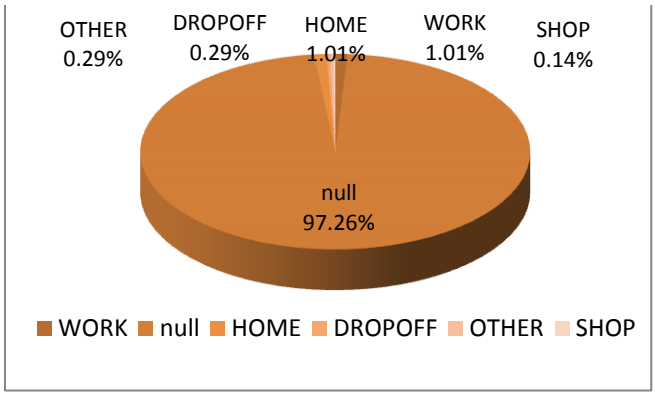

Figure 7. Purpose of trips for which transport mode was incorrectly forecasted.

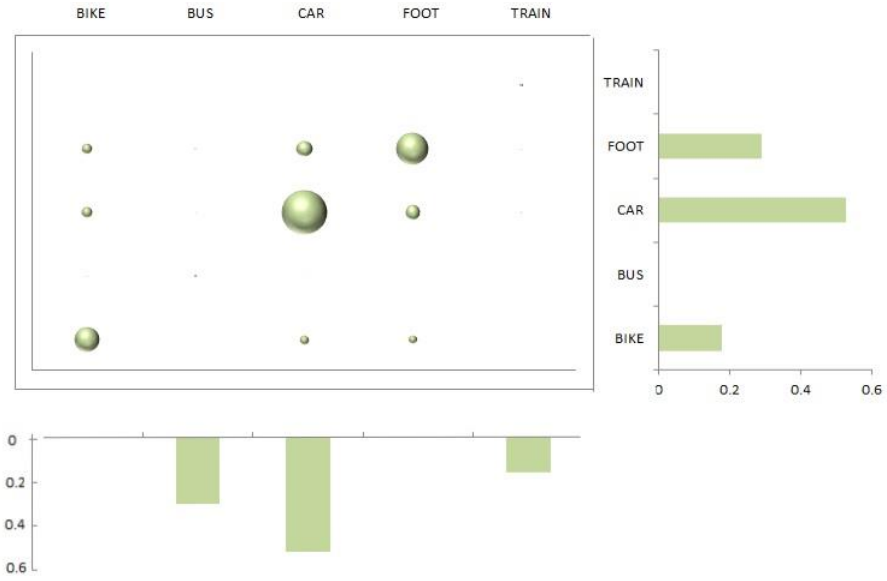

Figure 8. Confusion matrix.

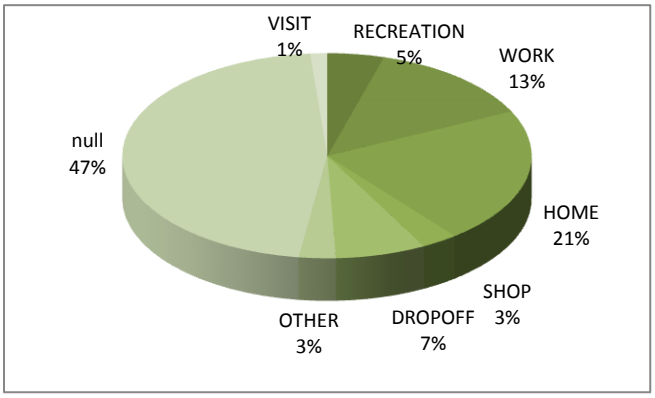

Figure 9. Purpose of trips for which transport mode was correctlly forecasted.

IV. DISCUSSION

Although the implemented SVM model has a quite high success rate (82%), still there is a place to improve forecasting results by additional extensions and considerations based on the gained insights. Firstly, since the model would give a good forecast in a bit more than eight out of ten trips, when implemented to provide pre-travel information, it could be useful to provide information on two transport modes that are most likely to be used. This way the model outcomes could be of higher relevance to the user and have outweighing role when user is considering more than one option. Furthermore, this way provision of such information could be also integrated into the city's transport management system and provided information can be coordinated with city's preferences. For example, if city wishes to promote new public transport line, or bike route, it can add additional weight to these options in the model so

 ISBN: 978-1-61208-510-4

that this information is provided to the user whenever feasible route using promoted options exists. In addition, there is a highest confusion between use of private car and active transport modes like bike and foot, which are the two ends of the sustainable mobility spectrum. It is worth to examining in future research in more details a context of trips for which confusion happened. Potential reasons include bad weather conditions when users are more prone to use private cars but also different trip purposes (e.g., one is more likely to use private car to drop-off multiple family members than for recreation). First reason can successfully be examined by fusing weather data and crowdsourced dataset. This insight car provide more descriptive context of the confusion occurrence. Second challenge is the familiarity of trip purpose. As forecasting time frame was an hour, this means that trip purpose should be known in advance. However, unless user indicates this information, a need to make a trip for certain reason should also be forecasted. This adds additional complexity to the model and can impact the success rate of the forecast. Therefore, it could be beneficial to investigate in more details complementarity of trip purpose forecasting models with transport mode forecasting and evaluate added value they give to each other.

Compared with results from literature [20], where gradient boosting trees were used to forecast transport mode one will use for the next trip, our results achieved with the support vector machines based model have around 10% higher success rate. This shows potential of support vector machines based model to be extended to incorporate other data sources and to be successfully implemented in order to support smart city mobility planning and managing process.

V. CONCLUSION

Support vector machines based model achieved success rate of 82% in forecasting transport mode one will use for the next trip. This shows high potential to implement such a model into smart city mobility system management and planning processes as it can result in development of more advanced pre-travel information service. Furthermore, gained insights already indicate potential future extensions of the model in order to ensure higher usability of the output results and improved relevance to the 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

[1] E. Tranos and D. Gertner, "Smart networked cities?" Innovation: The European Journal of Social Science Research. 25, pp. 175–190, 2012.

[2] P. Neirotti, A. de Marco, A. C. Cag, G. Mangano, and F. Scorrano. Current trends in Smart City initiatives: Some stylised facts. Cities, 38, pp. 25–36, 2014.

[3] I. Marsa-Maestre, M. A. Lopez-Carmona, J. R. Velasco, and I. A. Navarro, "Mobile Agents for Service Personalization in Smart Environments". Journal of Network and Computer Applications, 3, pp. 30–41, 2008.

[4] S. Beswick, Smart Cities in Europe: Enabling Innovation; Osborne Clarke: London, UK, 2014.

[5] G. J. A. Alonso and A. Rossi, New Trends for Smart Cities; ATOS: Bezons, France, 2011.

[6] L. Atzori, A. Iera, and G. Morabito, "The Internet of Things: A survey". Computer Networks, 54, pp. 2787–2805, 2010.

[7] T. Nam and A. Pardo, "Conceptualizing Smart City with Dimensions of Technology, People, and Institutions". In Proceedings of the 12th Annual International Conference on Digital Government Research, College Park, MD, USA, 12–15 June 2011.

[8] The Climate Group, ARUP, Accenture & The University of Nottingham. Information Marketplaces, The New Economics of Cities; The Climate Group, London, UK, 2011.

[9] G. C. Lazaroiu and M. Roscia, "Definition methodology for the smart cities model". Energy, 47, pp. 326–332, 2012.

[10] Y. Lu and Y. Liu, "Pervasive location acquisition technologies: Opportunities and challenges for geospatial studies". Computers, Environment and Urban Systems, 36, pp. 105–108, 2012.

[11] R. M. Pulselli, C. Ratti, and E. Tiezzi, "City out of Chaos: Social Patterns and Organization in Urban Systems". International Journal of Design & Nature and Ecodynamics, 1, pp. 125–134, 2006.

[12] J, Novak, R. Ahas, A. Aasa, and S. Silm, "Application of mobile phone location data in mapping of commuting patterns and functional regionalization: A pilot study of Estonia". Journal of Maps, 9, pp. 10–15, 2013.

[13] O. Järv, R. Ahas, and F. Witlox, "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, 2014.

[14] S. Iqbal, C. F. Choudhury, P. Wang, and M. C. González, "Development of origin-destination matrices using mobile phone call data". Transportation Research Part C: Emerging Technologies, 40, pp. 63–74, 2014.

[15] F. Liu, D. Janssens, J. Cui, Y. Wang, G. Wets, and M. Cools, "Building a validation measure for activity-based transportation models based on mobile phone data". Expert Systems with Applications, 41, pp. 6174–6189, 2014.

[16] Y. Yuan, M. Raubal, and Y. Liu, "Correlating mobile phone usage and travel behavior—A case study of Harbin, China". Computers, Environment and Urban Systems, 36, pp. 118–130, 2012.

[17] H. Bar-Gera, "Evaluation of a cellular phone-based system for measurements of traffic speeds and travel times: A case study from Israel". Transportation Research Part C: Emerging Technologies 15(6), pp. 380-391, 2007.

[18] O. Järv, R. Ahas, E., Saluveer, B. Derudder, and F. Witlox,

"Mobile Phones in a Traffic Flow: A Geographical Perspective to Evening Rush Hour Traffic Analysis Using Call Detail Records". PLoS ONE, 7(11), pp. e49171, 2012

[19] I. Semanjski and S. Gautama, Sensing Human Activity for Smart Cities' Mobility Management, InTech, 2016 (In Press)

[20] I. Semanjski and S. Gautama, "Smart City Mobility Application—Gradient Boosting Trees for Mobility Prediction and Analysis Based on Crowdsourced Data", Sensors, vol. 15, no. 7, pp. 15974-15987, 2015.

[21] I. Semanjski, A. Lopez Aguirre, J. De Mol and S. Gautama, "Policy 2.0 Platform for Mobile Sensing and Incentivized Targeted Shifts in Mobility Behavior", Sensors, vol. 16, no.7, pp. 1035-1053, 2016

[22] Ghent University, "Routecoach," Google Play, [Online]. Available: https://play.google.com/store/apps/details?id=com.move.routecoach. [Accessed 27 May 2016].

[23] New Integrated Smart Transport Options, "NISTO," [Online]. Available: http://www.nisto-project.eu. [Accessed 03 MarcN 2016].

[24] Official mobility statistics for Flanders, OVG, [Online] Available: http://www.mobielvlaanderen.be/ovg/. [Accessed 27 May 2016].

 ISBN: 978-1-61208-510-4