Computers, Environment and Urban Systems 66 (2017) 38–52

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/ceus

Spatial context mining approach for transport mode recognition from mobile sensed big data

Ivana Semanjski [a,*], Sidharta Gautama [a], Rein Ahas [b], Frank Witlox [c,b]

[a] Department of Telecommunications and Information Processing, Ghent University, Ghent, Belgium
[b] Department of Geography, University of Tartu, Tartu, Estonia
[c] Department of Geography, Ghent University, Ghent, Belgium

ARTICLE INFO

Article history:
Received 21 December 2016
Received in revised form 24 July 2017
Accepted 26 July 2017
Available online 10 August 2017

Keywords:
Transport mode recognition
Mobile sensed big data
Spatial awareness
Geographic information systems
Smart city
Support vector machines
Context mining
Urban data

ABSTRACT

Knowledge about what transport mode people use is important information of any mobility or travel behaviour research. With ubiquitous presence of smartphones, and its sensing possibilities, new opportunities to infer transport mode from movement data are appearing. In this paper we investigate the role of spatial context of human movements in inferring transport mode from mobile sensed data. For this we use data collected from more than 8000 participants over a period of four months, in combination with freely available geographical information. We develop a support vectors machines-based model to infer five transport modes and achieve success rate of 94%. The developed model is applicable across different mobile sensed data, as it is independent on the integration of additional sensors in the device itself. Furthermore, suggested approach is robust, as it strongly relies on pre-processed data, which makes it applicable for big data implementations in (smart) cities and other data-driven mobility platforms.

1. Introduction

To know the transport mode people use for their travel is a key element in any mobility study (Mc Fadden, 1978). Next to trip purpose, travel frequency and origin-destination pairs, transport mode information allows us to better understand people's travel behaviour (Bohte & Maat, 2009; Chen, Ma, Susilo, Liu, & Wang, 2016), manage traffic flows (Asakura, Tanabe, & Lee, 2000), and ensure a better transport and urban planning strategy (Magnanti & Wong, 1984). In general, transport modes can be seen as competing with or complementing one another in terms of cost, travel time, accessibility or comfort. It is not uncommon that mobility systems are compared, in regard to sustainability and efficiency of transport policy measures, based on achieved modal shifts (Banister, 2008; Chapman, 2007; Davison & Knowles, 2006) or accomplished modal splits of overall population in a defined area (Banister, 2008; Nabais, Negenborn, Carmona Benítez, & Ayala Botto, 2015; Steininger, Vogl, & Zettl, 1996; Tabuchi, 1993). On a more localized level information on the transport mode used forms a basis for development of location-based mobility services, and/or delivery of targeted mobility messages such as route information (Raper, Gartner, Karimi, & Rizos, 2007; Semanjski & Gautama, 2016a; Semanjski & Gautama, 2015).

Clearly, transport mode data collection is a key issue. Traditionally, this information is collected based on the questionnaires, travel diaries, and/or interviews. These data sources are also often referred to, in recent literature, as small data (Chen et al., 2016; Semanjski & Gautama, 2016b). Alternatively, big data based methods and techniques include extraction of transport mode information from Global Navigation Satellite Systems (GNSS), Bluetooth and/or mobile sensing techniques or by inferring it from mobile sensed big data (Semanjski & Gautama, 2016b; Toole et al., 2015; Vlahogianni, Park, & Van Lint, 2015). In this context, big data is often typified by the so-called 3Vs definition (Chen, Mao, & Liu, 2014; Semanjski et al, 2016; Witlox, 2015) where the three Vs stand for increase of volume (data scale becomes increasingly big when compared with small data), variety (data come in different formats, as structured, semi-structured and unstructured data) and velocity (data generation frequency strongly increases resulting in need for timely conduction of data collection-processing chain). In this paper, we explore potential to extract transport mode information from big data based on the trip's spatial context. Suggested approach relays on machine learning technique and complements existing approaches that mainly relay on variables that describe the moving object itself (as acceleration, speed etc.). Concretely, we aim to investigate the potential of implementing freely available geographical information and location information in order to infer and/or recognize transport modes from mobile sensed big data. To this end, we develop a new method, test this method on an extensive dataset collected over a four

* Corresponding author.
E-mail address: ivana.semanjski@ugent.be (I. Semanjski).

http://dx.doi.org/10.1016/j.compenvurbsys.2017.07.004
0198-9715/

months' timeframe involving more than 8000 individuals for the city of Leuven (Flanders, Belgium), and evaluate our approach. The paper is organized as follows. In Section 2, we present extensive literature review on the matter and position our approach in regard to the state of the art. In Section 3 we describe necessary data input. This relates to the collected mobile sensed data, and the spatial context awareness to determine the transport mode. In Section 4 we develop our model. The details of implemented model are given in the Section 5 and followed by the presentation of the results. The final sections include discussion and the conclusion remarks.

2. Literature review

As briefly mentioned in the introduction section, several methods and techniques exist that are used to extract information on the transport mode one utilizes for his/her travels: (i) traditional questionnaires, travel diaries, and/or interviews; (ii) with the help of GNSS, Bluetooth and/or mobile sensing techniques; or (iii) by inferring it from mobile sensed big data.

2.1. Surveys and interviews

Surveys and interviews are traditional and a rather straightforward way to collect data on transport mode people utilizes for their travels. They are conducted by means of paper or phone household surveys or interviews during which one is asked to record, or state, his or her travel behaviour on an average weekday. However, many studies (Ettema, Timmermans, & van Veghel, 1996; Stopher & Greaves, 2007) have shown that data collected in this manner deviated systematically from actual travel behaviour. These deviations include, among others, respondents' tendencies to underreport small trips (Itoh & Hato, 2013), car drivers to underestimate and public transport users to overestimate their travel times (Clifton & Muhs, 2012). In order to avoid these pitfalls, paper travel diaries were introduced (Stopher & Wilmot, 2000). Here, people are asked to systematically note their travel behaviour details with respect to travel times, transport modes, trip purposes, and frequencies. The data collection time span usually covers a period of one full typical week during a non-holiday seasons. Arentze et al. (2001) and Groves (2006) report that participants tend to postpone filling in these diaries which resulted in obtaining incomplete and inconsistent information. Quite often this reflects in forgetting to mention some smaller trips (e.g. walking to nearby restaurant during the lunch break), and rounding time and distances (Witlox, 2007) or having difficulties in defining the exact locations of places they have visited. As smaller trips were usually made by active transport modes, like cycling or walking (Declercq, Janssens, & Wets, 2013; Saelens, Sallis, & Frank, 2003), such data collection practice resulted in bias observed modal splits and further underpinned evolution of car-oriented transport planning.

2.2. GNSS and mobile sensing

Global navigation satellite systems opened new horizons in the collection of travel data (Feng & Timmermans, 2014; Wolf, Bricka, Ashby, & Gorugantua, 2004; Wolf, Guensler, & Bachman, 2001). By providing high resolution tracking data, GNSS collected data showed potential to overcome some of the disadvantages of the more traditional, abovementioned, approaches (Bohte & Maat, 2009; Bricka, Sen, Paleti, & Bhat, 2012; Montini, Prost, Schrammel, Rieser-Schüssler, & Axhausen, 2015). However, GNSS-logging exhibited a number of major restrictions as devices were typically installed in vehicles (Turner, Eisele, Benz, & Holdener, 1998). Hence, they only tracked a small portion of mobility behaviour (i.e., vehicle trips). Alternatively, when using portable handheld GNSS-devices, effort and discipline from the respondent to continuously carry the device with him/her are required, as forgetting the device results in unreported gaps in the trip data. Further developments in the field of GNSS chipsets enabled their integration in mobile phones allowing new possibilities for tracking. Today, smartphones have the same capabilities as the portable GNSS-device. However, carrying a smartphone has also become a habit and is therefore considered less of a burden, reducing the risk of non-reported trips. Furthermore, while vehicles were potentially shared among family members, this is less of a case for smartphones. Thus, this way sensed mobility behaviour corresponds more to ones' true travelling patterns.

We can distinguish three general ways mobile phone data are sensed for mobility studies:

First, we have call detail record (CDR) and network signalization data. This represents standardized data, collected by mobile network operators for billing purposes. Such data include records of all activities such as calls, SMSs, internet and data services where each record includes spatial and temporal parameters. Their applicability in the scope of mobility studies has been investigated for rush hour analysis (Bar-Gera, 2007; Järv, Ahas, Saluveer, Derudder, & Witlox, 2012), detection of variability in human activity spaces (González & Hidalgo, 2008; Hoteit, Secci, Sobolevsky, Ratti, & Pujolle, 2014; Järv, Ahas, & Witlox, 2014; Noulas, 2013; Williams, Thomas, Dunbar, Eagle, & Dobra, 2015; Xu et al., 2016), correlation of mobility behaviour with land use (Toole, Ulm, González, & Bauer, 2012), detection of origin-destination pairs and traffic zones (Alexander, Jiang, Murga, & Gonzalez, 2015; Dong et al., 2015; Iqbal, Choudhury, Wang, & Gonzalez, M. C., 2014; Toole et al., 2015). Kang et al. (2010) tackled complexity of such data pre-processing and related spatiotemporal analysis. They implemented geographical mapping and statistical analysis to gain deeper insight into ones' personal mobility patterns over different days and times of a day. Furthermore, Xu et al. (2015) studied a home-based approach to understand differences between human mobility patterns across diverse urban areas based on hierarchical clustering algorithm. Although CDR and network signalisation data are collected by all network operators, require no additional effort by users, no additional financial resources for their collection, cover wide areas and large populations their usage for mobility, and other, studies is still hindered by a number of privacy and regulatory issues as well as some technological, business related and methodological ones (Ahas et al., 2014; Calabrese, Ferrari, & Blondel, 2014; Seidl, Jankowski, & Tsou, 2016; Vij & Shankari, 2015). Among the problems mentioned, potentially the most significant ones are those related to location precision (limited to cellular network base station locations) and time resolution (limited to users' activity or regular network location updates dependable upon the type/generation of the network).

Second, we have so-called 'passive' mobile phone application based logging. This refers to the use of dedicated applications that run as a GNSS-based data logger in the background on the smartphone. Use of such sensed data is examined for the purposes of investigating individual mobility patterns (Calabrese, Diao, Di Lorenzo, Ferreira, & Ratti, 2013; Shin et al., 2015), speed analysis (Huss, Beekhuizen, Kromhout, & Vermeulen, 2014) or traffic monitoring (Herrera et al., 2010). The main advantage of this approach comes from higher spatial and temporal resolution of collected data than it is the case for mobile network call detail record, which are also considered to be 'passively' generated, as they also do not require any effort from the user. However, constantly active GNSS sensor tends to drain smartphone battery quite fast. This results in increased burden for users who, consequently, need to charge their phones more frequently. Compared to 'active' logging, the main advantage is that there is no need for interaction by respondents. That said, data collected this way require demanding data processing and interpretation efforts when compared to 'active' logging.

Third, and finally, we have 'active' and/or 'interactive' mobile phone application based logging. It represents the use of interactive mobile applications where respondents can report additional trip data as start of the trip or transport mode. In its essence, it can be considered as detailed mobile travel diary with the GNSS logging. Such reporting was, for instance, used to investigate the influence of carbon dioxide emissions

information on mode choice (Brazil & Caulfield, 2013) and, mostly, as ground truth for the development of supervised machine learning models in order to replace parts of traditional travel surveys (Nitsche, Widhalm, Breuss, Brändle, & Maurer, 2014; Nitsche, Widhalm, Breuss, & Maurer, 2012; Semanjski & Gautama, 2015). However, although 'active logging' requires manual intervention by the respondent this burden seems to be limited because the reporting is restricted to short entries at the very moment of departure and arrival. As a consequence, time and location of the departure and arrival can be more accurately detected, and there is no need for demanding data processing as, for example, splitting GNSS-based track on parts travelled by different modes (Bohte & Maat, 2009; Safi, Assemi, Mesbah, & Ferreira, 2016; Semanjski & Gautama, 2016b). Furthermore, the impact on the smartphone battery is limited as only small portions (when actual travel takes places) are being tracked and user has the control over the tracking process. Thus the user can choose not to log activity when smartphone battery is low (which anyway would not have been logged if battery would go flat).

Overall, all mobile sensed data collection approaches exhibit higher levels of spatial-temporal details and confidence in determining travel behaviour when compared with traditional approaches. Despite this fact, they are still not at a mature level to replace traditional approaches in a seamless way. The main reason for this is that required additional processing and interpretation is needed in order to extract information traditionally feed to transport planning models as information on transport mode or trip purpose. This is one of the main obstacles in implementing large potential of mobile sensed data for smart city mobility applications and transport planning needs.

2.3. Detection of transport mode from mobile sensed data

To overcome the above-mentioned deficiencies, different approaches exist that try to infer transport modes from mobile sensed data. These approaches are mainly based on GNSS (Bohte & Maat, 2009; Gong, Chen, Bialostozky, & Lawson, 2012; Huss et al., 2014), or accelerometer data (Bolbol, Cheng, Tsapakis, & Haworth, 2012; Hemminki, 2013; Hemminki, Nurmi, & Tarkoma, 2013; Manzoni, Maniloff, & Kloeckl, 2010; Shin et al., 2015; Wang, Chen, & Ma, 2010; Zhou, Yu, & Sullivan, 2016). Bohte and Maat (2009) use rule based approach to derive transport mode and trip purpose from GNSS data collected over period of one week. They achieve success rate of 70% for five transport modes and find train and public transport modes as the most challenging ones to distinguish, with success rates of 34 and 0% respectively. Gong et al. (2012) use similar rule based approach on a much smaller dataset, but include some of spatial descriptors into their model. This addition resulted in comparable success rate for train trips (36%), but the correct detection of public transport trips significantly increased (of up to 65%), resulting in overall success rate of 83% for five transport modes. Furthermore, Huss et al. (2014) showed that the same level of accuracy, when inferring the transport mode from GNSS data by rule based approach without spatial descriptors, can be achieved if one is inferring all motorized transport modes (bus, train and car) as one class. Manzoni et al. (2010) take advantage of additional sensors integrated into smartphones and implement decision trees based approach to differentiate between seven transport modes from data collected during one day from four persons. They achieve comparable success rate of 82%. Bolbol, Cheng, Tsapakis, and Chow (2012) use speed, acceleration, distance and changes in heading from accelerometer and GNSS sensors as input for support vector machines based model. They inferred six transport modes with success rate of 88% demonstrating the applicability of supervised machine learning based approach for transport mode recognition. Stenneth, Wolfson, Yu, Xu, and Morgan (2011) explored six individuals' trajectories over a period of three weeks. Next to speed and accelerometer data, they considered rail and bus stop proximity and, by doing so, showed that random forest based approach success rate increased from 76% to 93%. Similar approach was implemented by Rasmussen, Ingvardson, Halldórsdóttir, and Nielsen (2015), who used tracking data of 101 individuals, over five days, and fuzzy rule based logic. Xiao, Wang, Fu, and Wu (2017) derived around hundred accelerometer and speed based features and differed between six transport modes with success rate of 90%. Zhou et al. (2016) achieved the highest success rate by implementing chained random forest model to distinguish between motorized transport, biking and foot (walking and running) modes. They use data from 12 people's travel behaviour spanning over six days and correctly detect use of these three transport modes in 94% of cases. Integrated sensors, mainly accelerometers, are also used by Wang, Calabrese, Lorenzo, and Ratti (2010), Hemminki et al. (2013), Shin et al. (2015), Xiao, Juan, and Zhang (2015), Xiao et al. (2017) and Zhou et al. (2016). In addition, the use of cell tower information (Abdelaziz & Youssef, 2015) or call records (Wang et al., 2010) is also being explored. In many cases, studies report the implementation of combined approaches where the GNSS-data are used to improve accuracy of the accelerometer-based approach, or vice versa (Chen, Wang, Shen, Chen, & Zhao, 2013; Feng & Timmermans, 2013; Reddy, Burke, Estrin, Hansen, & Srivastava, 2008; Xia, Qiao, Jian, & Chang, 2014). The use of geographic information systems (GIS) for this purpose is reported less often (Biljecki, Ledoux, & Oosterom, 2013; Bohte & Maat, 2009; Rasmussen et al., 2015; Stenneth et al., 2011), whereas only few highlight the potential of big data and challenges it puts ahead (Chang, 2011; C. Chen et al., 2016; Witlox, 2015).

Table 1 gives an overview of the related literature with details on the use of GIS, size of the test data and number of transport modes that were inferred. Note that the option "standing still" is not considered to be a transport mode, as it does not represent transport facilities nor does it involve transporting people or goods, although some of the mentioned papers did differentiate it as a transport mode.

To summarise, on average, the published methods classify between three and six transport modes, and use around four indicators (Bolbol et al., 2012; Reddy et al., 2010, 2008). Furthermore, the detection of the transport mode mainly relies on the use of variables that describe the moving object itself, as speed and acceleration, implying that they give the highest indication of a transport mode (Bohte & Maat, 2009; Hemminki et al., 2013; Xia et al., 2014; Zhou et al., 2016). Here, the main challenge arises from similar speeds obtained by more than one transport mode (e.g., bike and pedestrians, or private car and public transport). This is only partially solved with the implementation of accelerometer data, but additional knowledge is still needed to increase the accuracy which is still mainly below 90%. Only a few studies have examined the contribution of GIS data to improve the accuracy (Biljecki et al., 2013; Bohte & Maat, 2009; Rasmussen et al., 2015; Stenneth et al., 2011), but these studies still relay on small datasets with very low variability in observed behaviour. Overall, all studies tested the proposed approaches on limited time span of collected data, ranging from four hours to three weeks (in addition to limited number of participants) failing to capture wide range of longitudinal time related (e.g. monthly) variations in travel behaviour. Furthermore, such short time ranges imply observed behaviour under similar environmental (e.g., weather) conditions and a small number of participants exhibits potential limitations in terms of transferability of the developed approaches on a wider population.

In order words, what is needed is a modelling approach that is able to counter these drawbacks. Our proposed method has three main advantages over previous approaches:

(1) our approach inverses the present paradigm where transport modes are inferred from variables that describe the moving object itself (as acceleration, speed etc.) to inferring transport mode from surrounding spatial contexts in which this movement takes place. This is particularly interesting as the suggested approach is applicable across different mobile and urban sensed data, and independent on the integration of additional sensors in the device itself;

Table 1
Related work on transport mode recognition from mobile sensed data.

Literature	Number of land transport modes	Use only GIS data	Use GIS data in combination with other	Duration of test data	Number of users	Accuracy
(Reddy et al., 2008)	5	No	No	240 min	6	90
(Bohte & Maat, 2009)	4	No	Yes	1 week	1104	70
(Wang et al., 2010)	5	No	No	12 h	7	70
(Reddy et al., 2010)	3	No	No	24 h	16	93
(Manzoni et al., 2010)	7	No	No	1 day	4	82
(Stenneth et al., 2011)	5	No	Yes	3 weeks	6	93
(Bolbol et al., 2012)	6	No	No	2 weeks	81	88
(Gong et al., 2012)	5	No	Yes	5 days	63	83
(Hemminki et al., 2013)	4	No	No	150 h	16	60–85
(Wang, Chen, & Chen, 2013)	5	No	No	N/A	5	90
(Biljecki et al., 2013)	7	No	Yes	1 week	1104	92
(Huss et al., 2014)	3	No	No	2 days	12	83
(Xia et al., 2014)	3	No	No	11 h	18	96
(Shin et al., 2015)	4	No	No	1 day	30	82
(Xiao et al., 2015)	5	No	No	5 days	202	86
(Rasmussen et al., 2015)	5	No	Yes	5 days	101	90
(Zhou et al., 2016)	3	No	No	6 days	12	94
(Xiao et al., 2017)	6	No	No	N/A	N/A	90

(2) our approach is tested on an extensive dataset of revealed travel movement; and,
(3) next to the location information our approach uses solely pre-processed and freely available geographical information to determine the transport mode. The latter makes our suggested approach robust and, in this sense, applicable for big data implementations in (smart) cities and other data-driven mobility platforms.

Thus, we aim to investigate the potential of implementing freely available geographical information and location information in order to infer and/or recognize transport modes from mobile sensed big data. To this end, we develop a new method, test this method on an extensive dataset collected over a four months' timeframe involving more than 8000 individuals for the city of Leuven (Flanders, Belgium), and evaluate our approach.

3. Data input and visualization

To develop our model, we use two data sets: (i) mobile sensed data that describes users' mobility behaviour; and (ii) publicly available geographical information that is used to define the spatial context in which this mobility behaviour took place. The city of Leuven (Flanders, Belgium) was used as focus area. Reason for this was a high density of collected trips and involved participants in this area. The city of Leuven (Fig. 1) is the capital of the province of Flemish-Brabant and located about 25 km east of Brussels (capital of Belgium). Brussels main airport is situated between two cities and connected with Leuven by road, rail and bike highway. The municipality itself comprises the historic city, surrounded by very busy ring road, and the former neighbouring municipalities (Heverlee, Kessel-Lo, a part of Korbeek-Lo, Wilsele and Wijgmaal) with total area of 57.51 km^2 (Leuven, 2016). Leuven has almost 100,000 inhabitants, and is home to one of the largest and oldest university in Belgium (Katholieke Universiteit Leuven) with more than 50,000 students (Leuven, 2016). Hence, our urban context is a very dynamic city which results in a lot of traffic and also related traffic congestion. It is characterised by a continuous growth in its public transport use (a fivefold increase in last 20 years) and a large share in cycling (17–20% modal share) (Declercq et al., 2013).

3.1. Mobile sensed data

Data on mobility behaviour is collected via an Android smartphone application (called Routecoach) developed at Ghent University, for the province of Flemish-Brabant, in the frame of the Interreg IVb NWE project 'New Integrated Smart Transport Options' (NISTO). The aim of NISTO was to develop an evaluation and planning toolkit for mobility projects which is applicable transnationally and can be adopted by planners. During the project lifespan (2013–2015), data on more than 150,000 trips was collected. A trip is defined as one-way movement

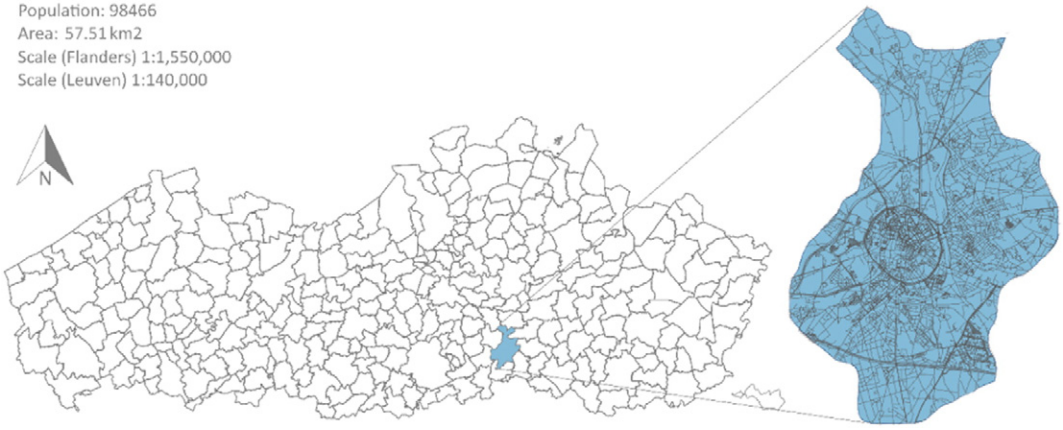

Fig. 1. The location of Leuven in Flanders.

Table 2
Sample – metadata.

Variable	Value
Users	8303
Trips	30,000
Transport modes	5
Time period	4 months
GNSS points	3,960,234
km	340,000
Median Points per trip	1243
Median Trips per user	8.7

between origin and destination points (e.g., home or work location) in the traffic network. The trip can be composed of smaller parts (called trip legs or segments) made by different transport modes. The Leuven data collection process happened between January and April 2015. In total, 8303 users actively participated by downloading the freely available application and collecting the data on one or more trips. In total more than 30,000 trips have been recorded leading to about 400,000 km of recorded data in driving, public transport, biking and walking. This data gives insight in different aspects of local and regional mobility: bike and car accessibility of the city, the observed walkability of the city centre, linked mobility etc. (Gautama, Van De Weghe, & De Maeyer, 2015). In order to start collecting data about a trip, the respondent first had to select between five pre-offered transport modes in the application (i.e., car, public transport, train, bike, or on foot) (s)he used for a given trip. The option "public transport" involves buses, trams and metro services, but in the case of Leuven it refers only to the use of the bus service as tram and metro services are not available. Reporting of transport mode changes during a trip was made possible by simple drag and drop option, from current transport mode to the new one, which minimized the user effort in recording timestamp, location and transport mode change. Also, in the application, the user needed to mark the end of the trip when reaching the destination. This way, for every recorded trip, transport mode and GNSS locations (with frequency of 1 Hz) were collected. Furthermore, all the participants had password protected access to their personal trip records where, through the user friendly GIS interface, they were able to perform data quality control and validate or correct any wrongly introduced trip information. In total, nearly 4 million GNSS location points were registered. The shortest trip had only 30 location points (corresponding to 58 m made by walking) and the longest 3047 points (equal to 802 km made by car) (Table 2). Most of the collected trips were made by bike (56%), followed by car (24%) and walking (11%) and the least by train (2%). Fig. **2** gives an impression of the density of the trips made by bike (in orange) and car (in blue). Car made routes clearly indicate the ring road that surrounds the city with major transport axes going from North to West of the area (European route E314) and radially around the main ring (national road N2 connecting Leuven to Brussels (North-West) on one side and Hasselt (North-East) on the other; N3 connecting from Brussels (South-West) area towards Walloon region (South-East); N264 on South, N19 and N26 on the North of the ring). Main rail lines tangents the ring on the East (rail station location) and bike routes highlight the city centre with pedestrian zone and connections towards university campuses on the South-West and South–East areas outside the main ring road.

3.2. Spatial context

The spatial context of the respondent's location is defined based on characteristics of objects situated in its surrounding. The awareness of this spatial context comprehends knowledge of these objects existence, purpose and proximity to the observing point of view. To be able to delineate the spatial context for collected trips we rely on freely available GIS data from OpenStreetMap (OSM). OSM is a prominent example of volunteered geographic information (Haklay, 2010; Jiang & Thill, 2015). In other words, data are provided on a voluntarily basis and, for our target area, many mobility-oriented information are included (e.g. locations of car sharing points, parking, train stations, railways, cycleways etc.). For orientation, there are 2 km of cycleways and 2 km of footways, as well as 14 km of other roads, per square km in Leuven. For each of these we defined its surrounding areas by creating 10 m, 30 m and 50 m buffers around them. Depending on in which of the buffers the GNSS point belonged to, and how many of them, the spatial context for that point and trip is defined. Thus, if we have railway and motorway one close to another, one GNSS point can be within, for example, 30 m of both motorway and railway. This would indicate that there is equal probability that the point belongs to trip made by car (motorway) or rail (railway), and none that the trip was made by bike as it is not allowed nor on motorway nor on railway. However, if we take a look at all points at trip level a clearer conclusion of the transport mode for the whole trip can be drown, as, for example, majority of

Fig. 2. Bike (orange) and car (blue) trips in the Leuven area (Flanders).

points can be within 10 or 30 m around the motorway (and minority around the railway), thus car transport mode is more likely to be utilized for that trip. Additional confirmation comes from taking a deeper look at trips' origin and destination points and if the trip both start and ends at known car park locations than this can be seen as confirmation of the assumption that the trip was made by car. Following this logic, we defined spatial context for the trips by taking a look at aggregated values for points at the trip level, expressed in percentages, and additionally, both trips' starting and ending locations. Table 3 gives full overview of predictor variables that were used in our model to define the spatial context of the mobile sensed trips and Fig. 3 visualizes this spatial context awareness principle where three so-called 'predictor variables' are noted in vicinity of the smartphone: an EV charging station (overall there are 14 EV charging stations in Leuven), a public transport stop (overall there are 652 public transport stops), and a bike highway (Leuven is connected with two bike highways, one going to the West towards Brussel and one going North towards Antwerp). Thus, for every trip, we analysed the spatial context of every observed location point, aggregated values at the trip level and trip's starting and ending point's spatial context.

Fig. 4 and Fig. 5 show examples of two mobile sensed trips, where one was made by bus transport and one by train, placed in the spatially aware context. For the clarity of visualization, the figures show only a part of spatial context related to the respective transport mode. Although, for example, the bus service trip starting and ending points are not located in close vicinity of recognised public transport stops it is still noticeable that most of the trip's locations are in vicinity of public transport lines. Similar is noticeable for the train trip example, where it can be seen that proximity of trips' locations to the railway defined context varies along the way. Potential reason for this could be that the GNSS accuracy is not constant or the railway context is defined based on the railway central line, whereas it could be that different tracks are closer, of further away, from this central line.

Table 3
Model predictor variables.

Predictor variables	Unit
Trip's points close to motorways (class <10 m; 10–30 m; >50 m)	[%]
Trip's points close to trunk roads (class <10 m; 10–30 m; >50 m)	[%]
Trip's points close to primary roads (class <10 m; 10–30 m; >50 m)	[%]
Trip's points close to secondary roads (class <10 m; 10–30 m; >50 m)	[%]
Trip's points close to tertiary roads (class <10 m; 10–30 m; >50 m)	[%]
Trip's points close to unclassified roads (class <10 m; 10–30 m; >50 m)	[%]
Trip's points close to residential, service or living roads (class <10 m; 10–30 m; >50 m)	[%]
Trip's points close to tracks, bridleways or paths (class <10 m; 10–30 m; >50 m)	[%]
Trip's points close to public transport lines (class <10 m; 10–30 m; >50 m)	[%]
Trip's points close to railway (class <10 m; 10–30 m; >50 m)	[%]
Trip's points close to cycleways and bike highways (class <10 m; 10–30 m; >50 m)	[%]
Trip start points in vicinity of bus station (class <10 m; 10–30 m; >50 m)	[Boolean]
Trip start points in vicinity of train station (class <10 m; 10–30 m; >50 m)	[Boolean]
Trip start points in vicinity of car parking, car wash, car sharing location, car repair, electric vehicle charging station (class <10 m; 10–30 m; >50 m)	[Boolean]
Trip start points in vicinity of bike parking (class <10 m; 10–30 m; >50 m)	[Boolean]
Trip end points in vicinity of bus station (class <10 m; 10–30 m; >50 m)	[Boolean]
Trip end points in vicinity of train station (class <10 m; 10–30 m; >50 m)	[Boolean]
Trip end points in vicinity of car parking, car wash, car sharing location, car repair, electric vehicle charging station (class <10 m; 10-30 m; >50 m)	[Boolean]
Trip end points in vicinity of bike parking (class <10 m; 10–30 m; >50 m)	[Boolean]

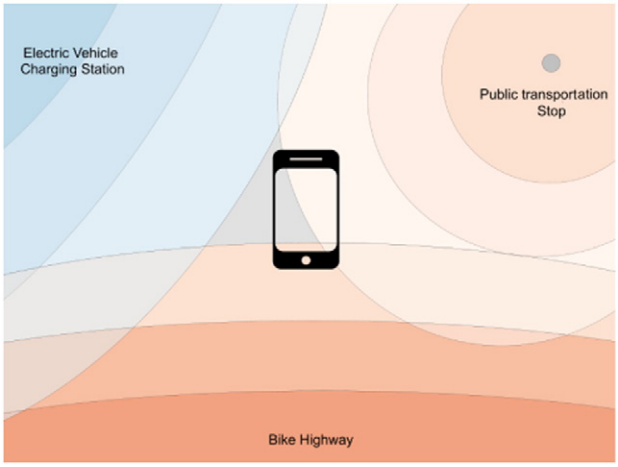

Fig. 3. Spatially aware context.

Next, it is our goal to develop a model that will be able, based on the sensed trips' location points and defined spatial context, to successfully infer utilized transport mode. For this, we adopted a support vector machines algorithm for which we describe the underlying principles in the following section.

4. Support vector machines classification

Support vector machines (SVMs) are supervised machine learning techniques widely used among data scientists for different purposes. The SVMs were first introduced in 1963 (Vapnik & Yakovlevich, 1963) and are still being actively developed in line with needs of different fields (Cortes & Vapnik, 1995; Hamel, 2011; McInerney, Stein, Rogers, & Jennings, 2013; Schlkopf & Smola, 2001). The standard SVMs are a non-probabilistic binary classifier that tries to separate between two classes by determining optimal separation hyperplane (Fig. 6). Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest points of any class (so-called margin), since the larger the margin the lower the generalization error of the classifier. Once the optimal hyperplane is defined, the decision to which class the new sample belongs to is done by simply checking on which side of the hyperplane the new sample is situated at.

Whereas the standard problem is stated in a finite dimensional space, it often happens that in that space the sets to be discriminated are not linearly separable. For this reason, the original finite dimensional space is mapped into a much higher dimensional space, presumably making the separation easier in that space. To achieve this, kernel functions ($K(X_i, X_j)$) are used.

Furthermore, as the SVMs are supervised machine learning technique, they require a training dataset which is used to train the predictor based on the labelled examples. Once the predictor is trained, and the optimal hyperplanes defined, another dataset with labelled examples is used – i.e. the test dataset. The test dataset is hence not involved in the training process itself and is considered to be an independent dataset with role to test the accuracy of the predictor. Once the SVMs model reached satisfactory results on both labelled datasets, it is considered to be applicable to further, unlabelled, examples. As such, the SVMs can be used for both classification and regression problems and are often considered to be among the best "off-the-shelf" supervised learning algorithm (Ng, 2016). Their applicability for transport planning related research has been shown by Semanjski (2015), Vlahogianni (2015) and Wang and Shi (2013) who implemented SVMs regression to forecast different continuous variables like speed or travel times. Furthermore, Cavar, Kavran, and Petrovic (2011) and Joo, Oh, Jeong, and

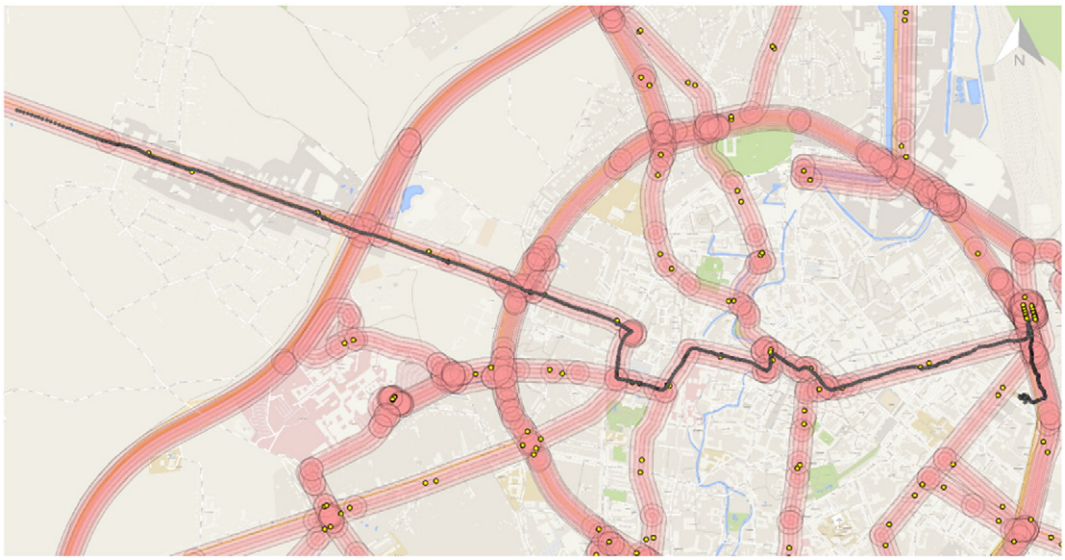

Fig. 4. Spatial visualization of public transport mode trip with highlighted areas around public transport lines (10, 30 and 50 m) and bus stops (yellow points).

Lee (2015) developed the SVMs classification based models to categorize urban roads depending on their traffic flow performance and bike use using observed GNSS traces. For these reasons, we explored the potential to develop the SVMs based model that can be trained to learn on user reported transport mode from labelled spatial context data (as defined in Table 3). Hence, in order to implement the SVMs based approach, we firstly needed to transform the transport mode inferring problem, which is a multiclass SVMs problem (not binary one), in form solvable by binary classifier, which SVMs are. For this purpose, we reduced the single multiclass problem into multiple binary classification problems by adopting a dummy variables approach. Here, we created dummy variables with case values as either 0 or 1 for each categorical variable and apply one vs. all classification. Thus, a categorical dependent variable consisting of five levels, say (A, B, C, D, E), is represented by a set of five dummy variables:

$$A:\{1\,0\,0\,0\,0\},B:\{0\,1\,0\,0\,0\},C:\{0\,0\,1\,0\,0\},D:\{0\,0\,0\,1\,0\},E:\{0\,0\,0\,0\,1\} \quad (1)$$

and, for a given example, is solvable by five binary classifiers. More details on multiclass classification can be found in literature (Galar, Fernandez, Barrenechea, Bustince, & Herrera, 2011; Hsu & Lin, 2002), as well as some practical examples in Anguita, Ghio, Oneto, Parra, and Reyes-ortiz (2012), Wu, Lee, and Yang (2008) and Semanjski and Gautama (2016a). Secondly, we divided the overall (labelled) dataset into two parts: the training dataset (Z_1) that included 75% of all the

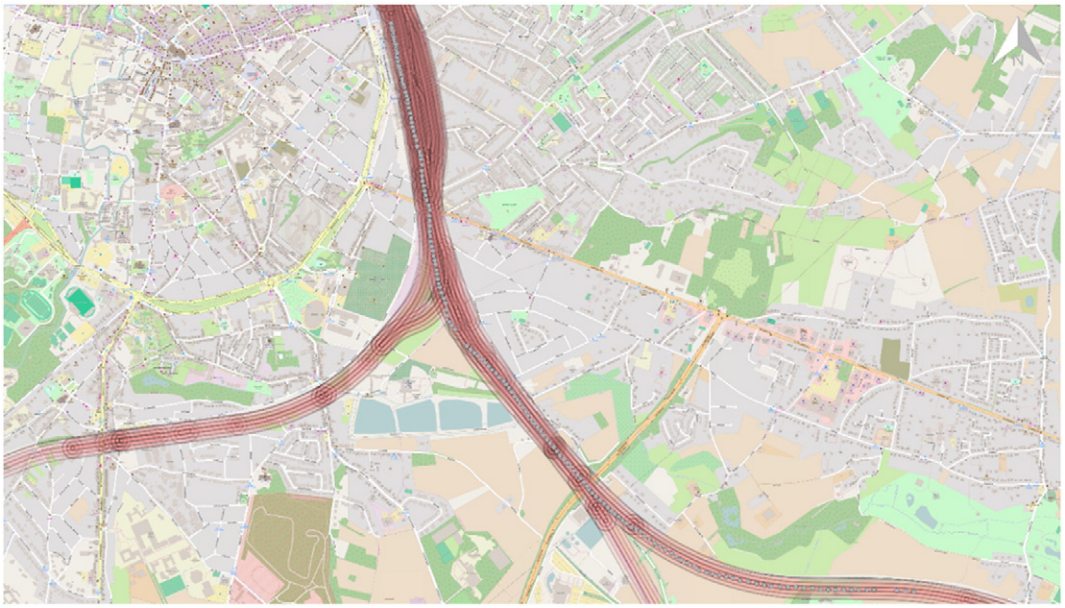

Fig. 5. Spatial visualization of train mode trip with highlighted areas around railway network (10, 30 and 50 m).

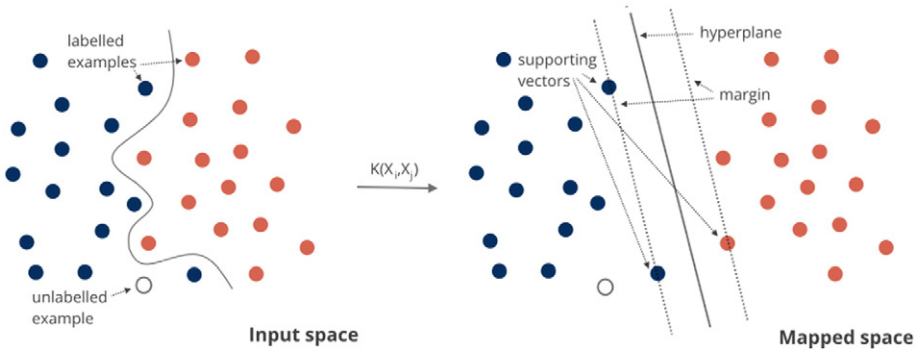

Fig. 6. Support vector machines basic principle.

data, and test dataset (Z_2) used to test the results of the training process, that contained the remaining 25% of the data. For training we adopted an error minimization function suggested by Vapnik (1998):

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

Subject to the constraints:

$$y_i(w^T\,\phi(x_i) + b) \geq 1 - \xi_i \tag{3}$$

$$\xi_i \geq 0 \tag{4}$$

where C is the capacity constant that controls the trade-off between errors of the SVMs on training data and margin maximization, w is the vector of separating hyperplanes' coefficients, b is a constant, and ξ_i represents parameters for handling non-separable data (inputs). The index i labels the N training cases ($y \in \pm 1$ represents the class labels and x_i represents the independent variables), and ϕ is a linear kernel function that transforms the inputs to the feature space:

$$K(X_i, X_j) = \phi(X_i)\cdot\phi(X_j) = X_i, X_j \tag{5}$$

Knowing that the selection of the capacity constant C (Eq. (2)) value is important to keep the training error small and in order to generalize well (Anguita & Oneto, 2011), we performed a two stage SVMs training. This means that in the first stage we introduced a grid search in range [1, 10] with increment 1 to find the best value of C. For this, we implemented a 10-fold cross-validation and found that the best value for capacity constant is 3. In the second training stage, the estimated value of capacity is applied to train an SVMs classifier $d(x_n)$ using the entire training sample. Following this training phase, the accuracy of the test data estimate (the proportion of cases in the test dataset which are misclassified by the classifier constructed from the training dataset) is calculated as follows:

$$R(d) = \frac{1}{N_2}\sum_{(x_n, j_n)\in Z_2} X(d(x_n) \neq j_n) \tag{6}$$

where X is the indicator function for which it is valid:

$X = 1$, if the statement $X(d) \neq j_n)$ is true.

$X = 0$, if the statement $X(d) \neq j_n)$ is false.

Once both, the training and test, phases are completed with success we can consider that optimal hyperplanes have been found and that the overall model is applicable for use with, new, unlabelled data.

Fig. 7 presents the conceptual SVMs based model for inferring transport mode from spatial contexts. Thus, firstly mobile sensed data need to be validated (labelled) by the user in order to obtain the ground truth on the transport mode used. As previously described, this process was performed directly in the app and/or, if needed, by means of a user friendly GIS web interface. Separately, the freely available GIS data are used and pre-processed to derive mobility oriented spatial context. These two insights are then joined so that each GNSS point, and consequently the whole trip, is placed into its spatially aware context. Aggregated values for predictor variables are obtained at trip level and joined by additional insight related to the trip's origin and destination locations. The datasets are then adjusted by mapping categorical variables and the validated dataset is divided into training and test parts. The training dataset is then feed to the SVMs learning process in order to define optimal hyperplanes. Once the two stage training process is

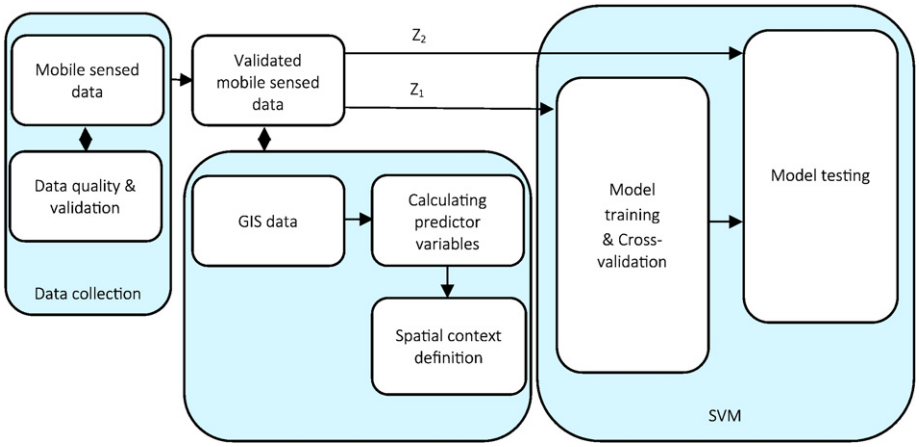

Fig. 7. Conceptual model.

completed, the obtained classifier is tested based on the independent (test) dataset.

5. Results

Based on the spatial context defined by predictor variables from Table 3, the adopted SVMs model was able to correctly infer transport mode utilized for trips in 94% of cases (Table 4). Regarding the complexity of the *n* dimensional space, overall there were 344 support vectors. The most support vectors (more than half) belonged to bike trips, and the least to train trips' observations (Table 5). This means that the separation of the mapped labelled samples was the most challenging for the transport mode belonging to the class bike, leaving the smallest margin space.

However, when inferring the transport mode, the confusion was the highest for the transport mode foot, misclassifying it as a bike in 18% of the cases. On the other end, no misclassifications occurred for transport modes bike and train (Fig. 8). Taking into the account that the least support vectors belonged to the train trips' observations (Table 5) and that no misclassification occurred for these trips (so, the train trips were not misclassified as made with another transport mode nor the trips made with other transport modes were misclassified as made by train) clearly indicates high potential of spatial context based approach when inferring train trips from mobile sensed data, which was so far one of the main challenges with the existing, moving object characteristics based, approaches (Bohte & Maat, 2009; Gong et al., 2012).

Overall, the highest success rate was achieved for transport modes bike and train, followed by car and public transport (Table 6), and the resulting modal split (Fig. 9) was just slightly changed in favour of bike (and disfavour of walking trips). It is interesting to note that out of 6% incorrectly recognised transport modes, the correct one was indicated as the second best choice in 61% cases, third best in 26% cases and was never the last one. This indicates potential to further improve the suggested approach.

Considering the high number of support vectors related to the detection of bike trips and confusion between walking trips and bike trips, potential reason why distinguishing between these two transport modes is challenging might be the fact that they often share the same urban space. In order to gain further understanding of confusions made, we take a deeper look on impact of shared infrastructure, land use characteristics and addition of new knowledge to the existing model in the next subchapters.

5.1. Shared infrastructure

Taking a look at the misclassifications one can also notice that some public transport trips (bus) were misclassified as car trips and some car trips as bike trips. Similarly, as with the misclassification of walking trips, the most obvious reason for this is the fact that the same space, and hence the spatial context, if often shared between more transport modes. For example, a trail in a park is likely to be used by both pedestrians and cyclists, and a road in the city between cars and public transport. For this reason, we investigated how different transport network elements are used by different transport modes. To do so, for every trip we analysed how many GNSS points were sensed at different types of network elements and if there is a correlation between usages of these network elements. Hence, we looked for mutual relationships between usage of different types or network elements within a trip based on the trip tracks data. For example, if one trip is sensed along the public transport line is it likely that it will also be observed at urban roads and not likely that it will be observed at motorway or along the nature park trail. Or if a trip is sensed along the cycleway, is it more likely that it will also be sensed in the residential area rather than along the railway. Consequently, if there is high correlation between usage of two types of road network, e.g. public transport line and urban road, then this can explain that there is likely to be confusion between two transport modes that are expected to predominantly use this infrastructure (car and public transport). To do so, we considered OSM based network elements explained in more details in Table 7.

Fig. 10 shows the results of the correlations analysis between the ways transport network elements were used for individual trips. One can easily notice a positive, and significant, correlation between usage of public transport network and primary and secondary roads which, for example, can explain confusions between observed transport modes bus and car. On the other end, trips that were sensed within a residential area are not likely to be observed at motorways, primary, secondary or tertiary roads. This is due to the fact that trips within residential areas are often short walking or cycling trips that are not likely to utilize (or even are not allow to) transport network infrastructure predominantly dedicated to fast motorized traffic. Once again, railways had low correlation with use of other network elements confirming the advantage of context based approach for inferring this transport mode.

5.2. Performance in urban and rural areas

To furthermore analyse the applicability of the spatial context based approach for inferring transport mode from mobile sensed data, we examined potential differences in performance within urban and rural areas. The motivation for this analysis lies in the impact of the built

Table 4
Classification accuracy in percentages.

Cross-validation accuracy	Overall
76.533	94

Table 5
SVM model summary.

Parameter	Value
Number of independents	54
SVM type	C-SVM
Kernel type	Linear
Number of SVs	344 (0 bounded)
Number of SVs (BIKE)	199
Number of SVs (BUS)	76
Number of SVs (CAR)	22
Number of SVs (FOOT)	39
Number of SVs (TRAIN)	8

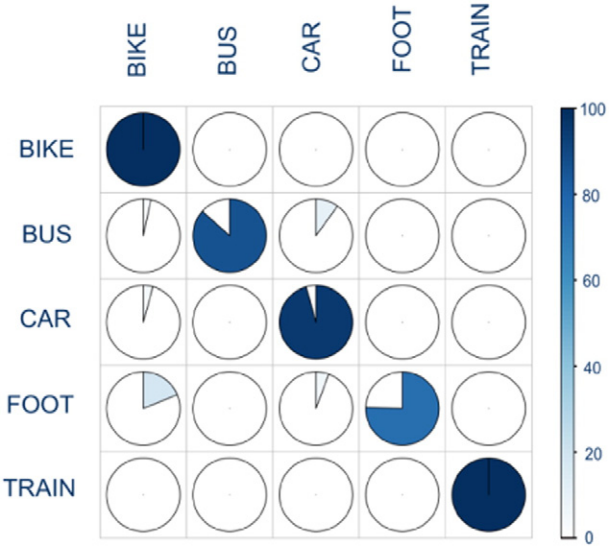

Fig. 8. Confusion matrix (each column represents percentage of trips made with the predicted transport mode while each row represents the percentage of trips made with an actual/validated transport mode).

Table 6
Classification results.

	Correct (%)	Incorrect (%)
BIKE	100.0	0.0
BUS	86.7	13.3
CAR	95.7	4.3
FOOT	75.5	24.5
TRAIN	100.0	0.0

environment on GNSS data precision and higher density of mobility oriented infrastructure in urban areas, both of which could result in different performances of the spatial context based approach. For this analysis we relied on land use data. Fig. 11 shows a map of the land use of the Leuven area classified in two categories. We distinguish urban as containing predominantly build up area with commercial and housing activities, and rural relates to areas without buildings such as parks, farmyards and fields. Firstly, we mapped all the GNSS points with this land use data, and 52% were located in, so-called, rural area, while 48% were in urban area. Fig. 12 and Fig. 13 illustrate the modal split for these two areas based on the validated data from users. The modal splits slightly differ, as in urban areas the active transport modes, such as cycling and walking, are slightly more represented, while major highways pass outside of this urban area, and thus car and train trips are more often sensed in rural areas.

Secondly, we mapped the GNSS points and their spatial context in a higher dimensional space with previously defined hyperplanes. Considering confusion matrixes, when the urban and rural areas of Leuven are treated separately, a slightly higher success rate is achieved for rural areas (Fig. 14 and Fig. 15) where the overall success rate was 97%, while in urban area it was 93%. In urban areas, most of the confusion occurred between transport modes car and bike and bike and walking, while in rural areas confusions were noticeably smaller. However, in both areas, the achieved results are seen as satisfactory, hence the approach yields good results regarding the different characteristics of urban and rural areas.

5.3. Adding a speed to the spatial context

Next to the analysis of the transport network relevance and the rural and urban area characteristics, we added a new (not spatial) predictor to the model – speed. The motivation for this exercise was to see can, and how much, non-spatial context related features add to the spatial context based approach for inferring transport mode from mobile sensed data. Hence, together with the predictor variables defined in Table 3, we included an additional continuous variable that describes the rate of change of moving object position expressed in kilometres per hour. This addition required a remapping of the labelled examples as a new dimension is added to the space and, respectively, calculation of new hyperplanes. Here, the same error minimization function (Eq. (2)) and kernel function (Eq. (5)) are applied as previously, and the first SVMs training stage yield the same best value for C. Overall, the model had 346 supporting vectors which is seen as comparable with the previous results, reflecting similar complexity when separating mapped labelled samples belonging to different transport modes (Table 8). Furthermore, overall success rate increased to 95.2% implying that speed can slightly contribute to the spatially aware context.

Table 7
Transport network types.

Transport network type	Description	OSM type
Motorways	A restricted access public road specifically used for fast traffic. It is not open to pedestrians, cyclists, mopeds and farm vehicles. Usually it has two or more running lanes plus emergency hard shoulder	'motorway' and 'motorway_link
Trunk roads	The most important roads in a country's system that aren't motorways.	'trunk' and 'trunk_link'
Primary roads	The next most important roads in a country's system. They often link larger towns.	'primary' and 'primary_link'
Secondary roads	The next most important roads in a country's system. They often link towns.	'secondary' and 'secondary_link'
Tertiary roads	The next most important roads in a country's system. Often link smaller towns and villages.	'tertiary' and 'tertiary_link'
Unclassified roads	The least most important through roads in a country's system. For example, minor roads of a lower classification than tertiary, but which serve a purpose other than access to properties. Often link villages and hamlets.	'unclassified' and 'unclassified_link'
Residential roads	Roads which serve as an access to housing or an industrial estate, without function of connecting settlements. Often lined with housing and streets where pedestrians have legal priority over cars, speeds are kept very low and where children are allowed to play on the street.	'residential', 'residential_link', 'service', 'living' and 'living_link'
Nature	Roads for mostly agricultural or forestry uses, tracks for walking and gravel paths, tracks for horses. Often rough with unpaved surfaces.	'track', 'track_link', 'bridelway', 'bridelway_link', 'path', 'path_link'
Public transport line	Way on which the bus travels on.	'public_transport', 'public_transport_link'
Railways	Metal rail lines generally assumed to be primarily for use by passenger trains.	'OSM type railway' and 'railway_link'

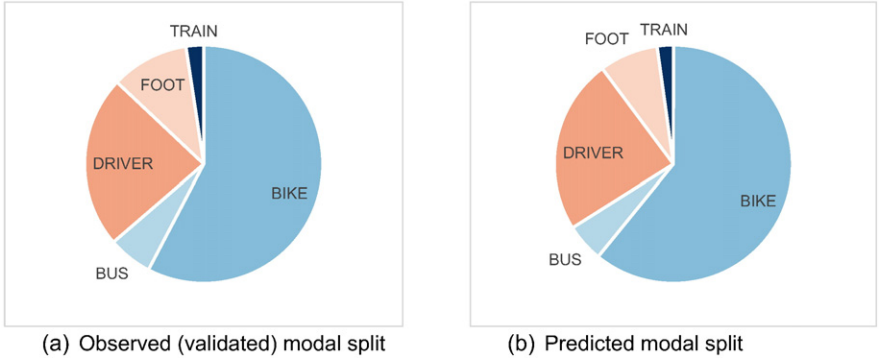

Fig. 9. Modal split - observed and predicted results.

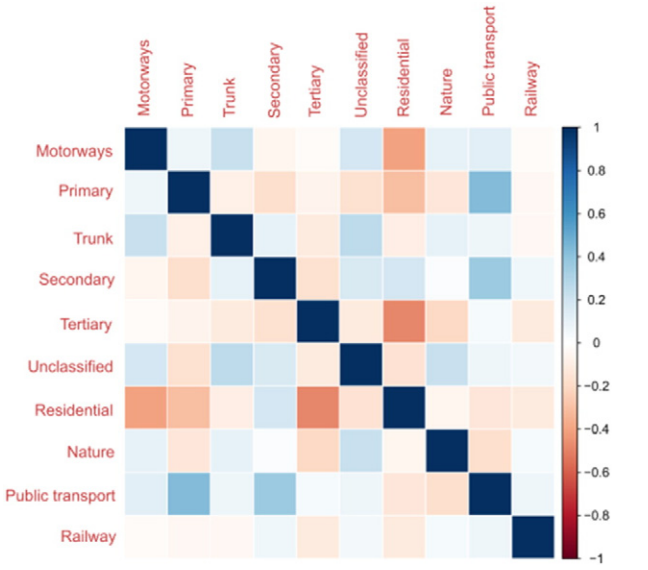

Fig. 10. Correlations between the ways infrastructure is used for individual trips.

Considering the confusion matrix (Fig. 16), now some confusion happened between the transport modes bike and walking. However, a good separation of the train transport mode still remained. The confusion level for transport mode walking reduced, implying highest improvement for this transport mode now that speed is added to the model. However, other transport modes (bus, bike and car) got now misclassified as walking, although these values are rather low (below 6%). Also, separation between transport modes walking and bike improved a little bit as now less trips are misclassified as bike, instead as walking trips.

6. Discussion and conclusions

We have investigated the likelihood to infer transport modes for different trips based on the smartphone location of the trip-maker and

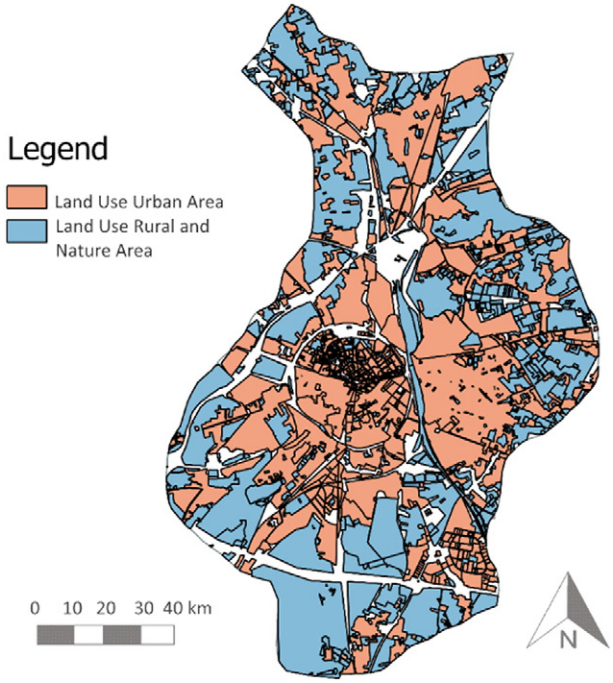

Fig. 11. Map of land use in Leuven area.

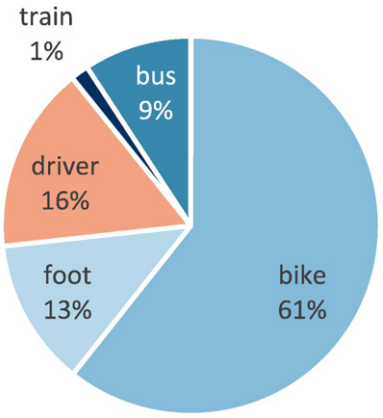

Fig. 12. Urban modal split.

freely available GIS data (OpenStreetMap). The GIS data were used to provide the spatial context of the trip-maker's movements, and based on this, the transport mode was recognised correctly in 94% of the trips utilizing one of five different transport modes and reported by more than 8000 users in the area of Leuven, Flanders. Compared to previous attempts and related papers (summarized in Table 1) higher success rates are reported only in one case where only three transport modes were inferred based on the accelerometer and GNSS data collected from just 18 users over a time span of 11 h. Our approach is tested on a far bigger and more diverse test set. In addition, our approach requires much less data to be transferred by smartphone (just location information) compared to other approaches, making the used approach more cost-beneficial considering real-time data transfer by the users' smartphones, and more energy efficient considering users' smartphone battery usage (Hemminki et al., 2013; Xia et al., 2014). Furthermore, as the transport mode detection was tested for the trips which had 30, or more, GNSS locations, for real-time implementations, the first results could be obtained quite fast (after the first 30 points are collected) where successive GNSS points could only improve initial results. The latter presents additional advantage over existing approaches that require a whole trip to be completed in order to determine transport mode one was using. In addition, post-processing of trip data is minimized as map-matching is not needed (the approach integrates raw GNSS points as input). This is highly relevant for big data applications as promptly and timely conducted data collection-processing chain maximizes the value of data and often differentiates between, simply, large data amounts and big data itself (Chen et al., 2014; Provost & Fawcett, 2013). Furthermore, our approach shows promising improvements when inferring train and public transport modes, which were identified

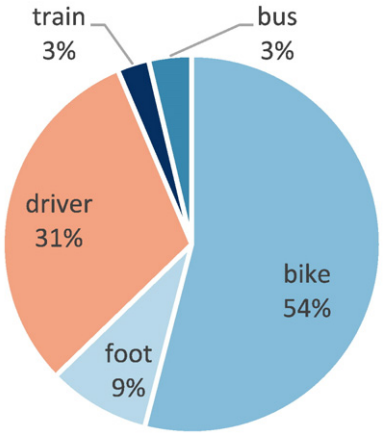

Fig. 13. Rural modal split.

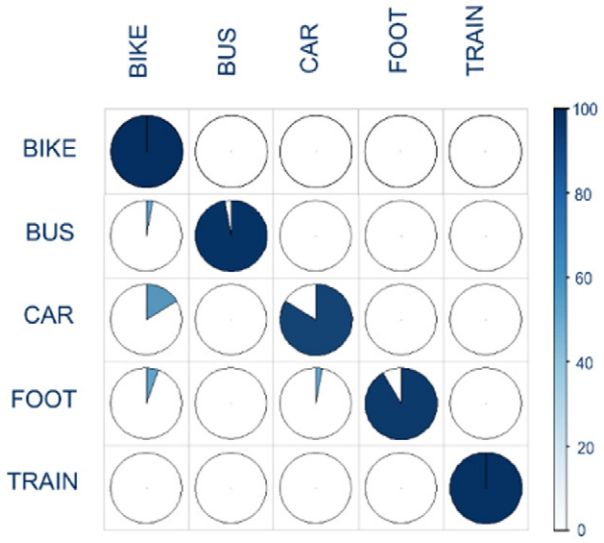

Fig. 14. Confusion matrix for urban area.

in the existing literature (Bohte & Maat, 2009; Gong et al., 2012) as the most challenging ones.

In more details, our approach successfully inversed the present paradigm where transport modes are inferred from variables that describe the moving object itself (as acceleration, speed etc.) to doing so from surrounding spatial contexts in which the movement takes place. This exhibits one of major advantages of the suggested approach which is its applicability across different mobile sensed data, independent on the integration of additional sensors in the device itself, and thus resilient towards high heterogeneity among sensing devices. In more details, the suggested approach can easily be extended to integrate other location aware urban sensed data to infer transport mode. One example of this would be Bluetooth sensors placed around the city that register active Bluetooth devices (as car keys, tablets etc.). By knowing the spatial context of the placed sensors, and considering observations of same device's ID across the city, one could extend suggested approach to infer used transport mode that best describes observed Bluetooth device movements. The latter makes suggested approach robust and, in this sense, applicable for big data implementations in (smart) cities and other data-driven mobility platforms.

Table 8
SVM model summary.

Parameter	Value
Number of independents	55
SVM type	C-SVM
Kernel type	Linear
Number of SVs	346 (0 bounded)
Number of SVs (BIKE)	199
Number of SVs (BUS)	78
Number of SVs (CAR)	22
Number of SVs (FOOT)	39
Number of SVs (TRAIN)	8

As a potential limitation, of using free GIS data for transport mode detection, data availability and their validity could be considered since those data are provided on a volunteering basis and their coverage varies among different countries and areas. This seems to be inherent to all citizen science and social sensing based approaches (Liu et al., 2015). As a hypothetical example of such limitation, unpredicted/unreported changes (disturbances) in transport network could result in a higher level of ambiguity in detecting between transport modes for real-time applications. However, existing statistics (Neis & Zielstra, 2014; Zhao, Jia, Qin, Shana, & Jiao, 2015) show an exponential growth of freely available GIS data (with increased accuracy and spatial coverage). This paves the way for applications, like ours, that rely on such data sources. Furthermore, a growing number of municipalities and communities is building its own, validated, open GIS datasets (Liu et al., 2015; Pan, Tian, Liu, Gu, & Hua, 2016; Thorsby, Stowers, Wolslegel, & Tumbuan, 2017). It is not uncommon that these datasets integrate timely updates on traffic conditions (e.g. information about road closures or similar). Availability of such data could only positively impact previously mentioned challenges related to the higher level of ambiguity for real-time applications. In addition, the main challenge of our approach comes from the fact that different transport modes often share same space and thus the same spatial context. This is a common case in many real-life situations as, for example, when public transport (bus) shares a line with private cars or the fact that pedestrians and cyclists both use a city square. Adding additional intelligence about moving object (as information on speed) improved achieved results slightly, but still it remained challenging to completely distinguish between all transport as, when sharing the same space, moving objects tend to adjust their behaviour to achieve synchronizes movements. An example

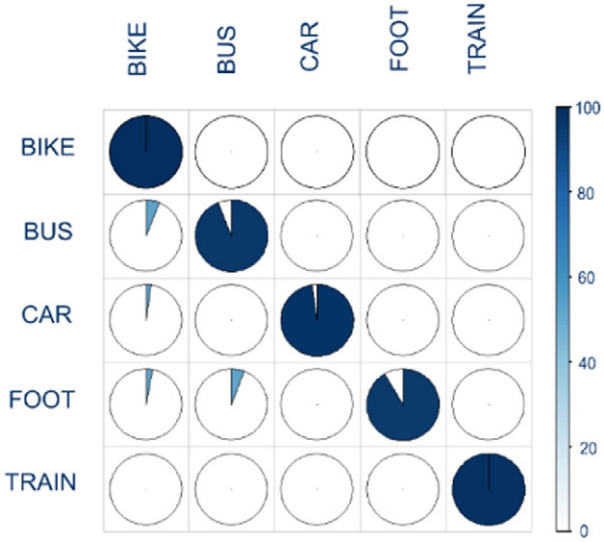

Fig. 15. Confusion matrix for rural area.

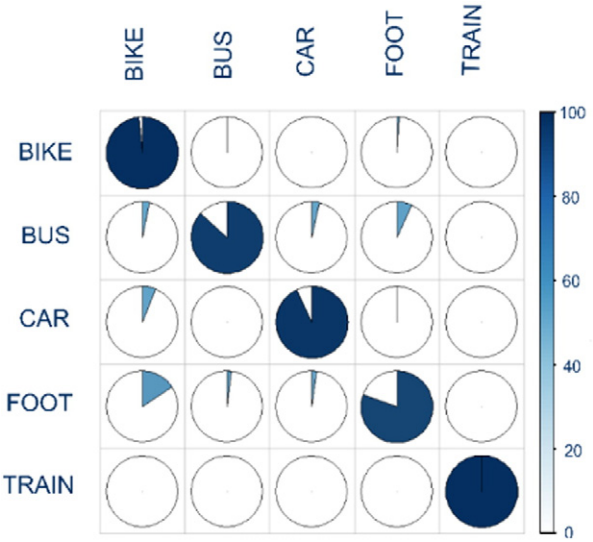

Fig. 16. Confusion matrix for speed refined spatial context based model.

of this could be a cyclist who will slow down to manoeuvre among the crowd and the fact that the speed limit will be the same, or similar, for public transport and private cars in the build-up area. However, there is potential in additional pre-processing of spatial data as encoding of improbable and impossible usage of a space. Therefore, by focusing, not on where it is expected to observe certain transport mode (e.g. bike on a bike path) but, on where it is highly unlikely to do so (e.g. trip made by foot is unlikely to have parts made along a motorway or train trip along the footpath in park or private car along streets allowed only for public transport). In this sense, joining the information on transport mode that is likely to be utilized for the trip and penalising the recognition of transport mode that is, from this context, unlikely to be utilized, has potential to further improve inferring results. Future research is needed on these hypotheses and the impact of newly added intelligence on the success rate of transport mode recognition.

Eventually, our approach is the first to successfully address some of the key issues recognised in literature (Chang, 2011; Witlox, 2015) related to the implementation of the big data for recognition of transport mode considering data size and power management of required data transfer. In this regards, our model it seen as easily implementable for a growing number of smart city mobility applications ranging from crowdsourced based mobility management to location based mobility notifications which are targeting at users of specific transport modes.

Acknowledgments

This research is funded by INTERREG North-West Europe (grant number 41W08913W) project New Integrated Smart Transport Options (NISTO), the Flemish government agency for Innovation by Science and Technology and the Flemish Institute for Mobility.

References

Abdelaziz, A. M., & Youssef, M. (2015). The diversity and scale matter: ubiquitous transportation mode detection using single cell tower information. *IEEE Vehicular Technology Conference*, *2015*(0), 2015–2017. http://dx.doi.org/10.1109/VTCSpring.2015.7146124.

Ahas, R., Armoogum, J., Esko, S., Ilves, M., Karus, E., Madre, J. -L., ... Tiru, M. (2014). *Feasibility study on the use of mobile positioning data for tourism statistics - consolidated report. http://dx.doi.org/10.2785/55051*.

Alexander, L., Jiang, S., Murga, M., & Gonzalez, M. C. (2015). Origin–destination trips by purpose and time of day inferred from mobile phone data. *Transportation Research Part C: Emerging Technologies*, *58*, 240–250. http://dx.doi.org/10.1016/j.trc.2015.02.018.

Anguita, D., Ghio, A., Oneto, L., Parra, X., & Reyes-ortiz, J. L. (2012). *Human activity recognition on smartphones using a multiclass hardware-friendly support vector machine* (pp. 216–223)International Workshop on Ambient Assisted Living IWAAL 2012. Springer.

Anguita, D., & Oneto, L. (2011). *In sample model selection for support vector machines*The 2011 International Joint Conference on Neural Networks (p. 2011). San Jose, CA.: IEEE.

Arentze, T., Dijst, M., Dugundji, E., Maat, K., Timmermans, H., Kapoen, L., ... Goodall, N. (2001). New activity diary format: Design and limited empirical evidence Record: Journal of the Transportation Research Board article/chapter tools search crashes. *Transportation Research Record: Journal of the Transportation Research Board*, *1*(1786), 79–88. http://dx.doi.org/10.3141/1768-10.

Asakura, Y., Tanabe, J., & Lee, Y. H. (2000). *Characteristics of positioning data for monitoring travel behaviour. In proceedings of the 7th world congress on intelligent systems. Tokio, Japan*.

Banister, D. (2008). The sustainable mobility paradigm. *Transport Policy*, *15*(2), 73–80. http://dx.doi.org/10.1016/j.tranpol.2007.10.005.

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), 380–391. http://dx.doi.org/10.1016/j.trc.2007.06.003.

Biljecki, F., Ledoux, H., & Oosterom, P. V. (2013). Transportation mode-based segmentation and classification of movement trajectories ∗. *International Journal of Geographical Information Science*, *27*(2), 385–407.

Bohte, W., & Maat, K. (2009). Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: A large-scale application in the Netherlands. *Transportation Research Part C: Emerging Technologies*, *17*(3), 285–297. http://dx.doi.org/10.1016/j.trc.2008.11.004.

Bolbol, A., Cheng, T., Tsapakis, I., & Chow, A. (2012). Sample size calculation for studying transportation modes from GPS data. *Procedia - Social and Behavioral Sciences*, *48*, 3040–3050. http://dx.doi.org/10.1016/j.sbspro.2012.06.1271.

Bolbol, A., Cheng, T., Tsapakis, I., & Haworth, J. (2012). Inferring hybrid transportation modes from sparse GPS data using a moving window SVM classification. *Computers, Environment and Urban Systems*, *36*(6), 526–537. http://dx.doi.org/10.1016/j.compenvurbsys.2012.06.001.

Brazil, W., & Caulfield, B. (2013). Does green make a difference: The potential role of smartphone technology in transport behaviour. *Transportation Research Part C: Emerging Technologies*, *37*, 93–101. http://dx.doi.org/10.1016/j.trc.2013.09.016.

Bricka, S. G., Sen, S., Paleti, R., & Bhat, C. R. (2012). An analysis of the factors influencing differences in survey-reported and GPS-recorded trips. *Transportation Research Part C: Emerging Technologies*, *21*(1), 67–88. http://dx.doi.org/10.1016/j.trc.2011.09.005.

Calabrese, F., Diao, M., Di Lorenzo, G., Ferreira, J., & Ratti, C. (2013). Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. *Transportation Research Part C: Emerging Technologies*, *26*, 301–313. http://dx.doi.org/10.1016/j.trc.2012.09.009.

Calabrese, F., Ferrari, L., & Blondel, V. D. (2014). Urban sensing using mobile phone network data: A survey of research. *ACM Computing Surveys*, *47*(2), 1–20. http://dx.doi.org/10.1145/2655691.

Cavar, I., Kavran, Z., & Petrovic, M. (2011). Hybrid approach for urban roads classification based on GPS tracks and road subsegment data. *PROMET – Traffic &Transportation*, *23*(4), 286–289.

Chang, E. Y. (2011). *Context-aware Computing : Opportunities and open issues, 3–4*.

Chapman, L. (2007). Transport and climate change: A review. *Journal of Transport Geography*, *15*(5), 354–367. http://dx.doi.org/10.1016/j.jtrangeo.2006.11.008.

Chen, C., Ma, J., Susilo, Y., Liu, Y., & Wang, M. (2016). The promises of big data and small data for travel behavior (aka human mobility) analysis. *Transportation Research Part C: Emerging Technologies*, *68*, 285–299. http://dx.doi.org/10.1016/j.trc.2016.04.005.

Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. *Mobile Networks and Applications*, *19*(2), 171–209. http://dx.doi.org/10.1007/s11036-013-0489-0.

Chen, Z., Wang, S., Shen, Z., Chen, Y., & Zhao, Z. (2013). *Online sequential ELM based transfer learning for transportation mode recognitionProceedings of the 2013 IEEE Conference on Cybernetics and Intelligent Systems, CIS 20130*. (pp. 78–83). http://dx.doi.org/10.1109/ICCIS.2013.6751582.

Clifton, K., & Muhs, C. (2012). Capturing and representing multimodal trips in travel surveys. *Transportation Research Record: Journal of the Transportation Research Board*, *2285*, 74–83. http://dx.doi.org/10.3141/2285-09.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297.

Davison, L. J., & Knowles, R. D. (2006). Bus quality partnerships, modal shift and traffic decongestion. *Journal of Transport Geography*, *14*(3), 177–194. http://dx.doi.org/10.1016/j.jtrangeo.2005.06.008.

Declercq, K., Janssens, D., & Wets, G. (2013). *Onderzoek Verplaatsingsgedrag*. (Diepenbeek).

Dong, H., Wu, M., Ding, X., Chu, L., Jia, L., Qin, Y., & Zhou, X. (2015). Traffic zone division based on big data from mobile phone base stations. *Transportation Research Part C: Emerging Technologies*, *58*, 278–291. http://dx.doi.org/10.1016/j.trc.2015.06.007.

Ettema, D., Timmermans, H., & van Veghel, L. (1996). *Effects of data collection methods in travel and activity research*. Eindhoven, Netherlands: European Institute of Retailing and Service Studies, 2000.

Feng, T., & Timmermans, H. J. P. (2013). Transportation mode recognition using GPS and accelerometer data. *Transportation Research Part C: Emerging Technologies*, *37*, 118–130. http://dx.doi.org/10.1016/j.trc.2013.09.014.

Feng, T., & Timmermans, H. J. P. (2014). Extracting activity-travel diaries from GPS data: Towards integrated semi-automatic imputation. *Procedia Environmental Sciences*, *22*, 178–185. http://dx.doi.org/10.1016/j.proenv.2014.11.018.

Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., & Herrera, F. (2011). An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes. *Pattern Recognition Journal*, *44*(2011), 1761–1776. http://dx.doi.org/10.1016/j.patcog.2011.01.017.

Gautama, S., Van De Weghe, N., & De Maeyer, P. (2015). From big data to smart citizens. In L. Boelens, D. Lauwers, & F. Witlox (Eds.), *A new policy and research agenda on mobility in horizontal metropolises* (pp. 123–138). Groningen: InPlanning.

Gong, H., Chen, C., Bialostozky, E., & Lawson, C. T. (2012). A GPS/GIS method for travel mode detection in New York City. *Computers, Environment and Urban Systems*, *36*(2), 131–139. http://dx.doi.org/10.1016/j.compenvurbsys.2011.05.003.

González, M. C., & Hidalgo, C. a. (2008). Understanding individual human mobility patterns. *Nature*, *782*(June 2008), 1–12. http://dx.doi.org/10.1038/nature06958.

Groves, R. (2006). Nonresponse rates and nonresponse bias in household surveys: What do we know about the linkage between nonresponse rates and nonresponse bias? *Public Opinion Quarterly*, *70*(5), 646–675. http://dx.doi.org/10.1093/poq/nfl033.

Haklay, M. (2010). How good is volunteered geographical information? A comparative study of OpenStreetMap and ordnance survey datasets. *Environment and Planning B: Planning and Design*, *37*(4), 682–703. http://dx.doi.org/10.1068/b35097.

Hamel, L. H. (2011). *Knowledge discovery with support vector machines*. Hoboken: Wiley-Interscience.

Hemminki, S. (2013). Transportation behavior sensing using smartphones. *International Joint Conference on Pervasive and Ubiquitous Computing. Zurich, Switzerland*.

Hemminki, S., Nurmi, P., & Tarkoma, S. (2013). *Accelerometer-based transportation mode detection on smartphones categories and subject descriptors*Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems. Rome, Italy.

Herrera, J. C., Work, D. B., Herring, R., Ban, X., Jacobson, Q., & Bayen, A. M. (2010). Evaluation of traffic data obtained via GPS-enabled mobile phones: The mobile century field experiment. *Transportation Research Part C: Emerging Technologies*, *18*(4), 568–583. http://dx.doi.org/10.1016/j.trc.2009.10.006.

Hoteit, S., Secci, S., Sobolevsky, S., Ratti, C., & Pujolle, G. (2014). Estimating human trajectories and hotspots through mobile phone data. *Computer Networks*, *64*, 296–307. http://dx.doi.org/10.1016/j.comnet.2014.02.011.

Hsu, C. -W., & Lin, C. -J. (2002). A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks*, *13*(2), 415–425. http://dx.doi.org/10.1109/72.991427.

Huss, A., Beekhuizen, J., Kromhout, H., & Vermeulen, R. (2014). Using GPS-derived speed patterns for recognition of transport modes in adults. *International Journal of Health Geographics*, *13*(1), 40. http://dx.doi.org/10.1186/1476-072X-13-40.

Iqbal, M. S., Choudhury, C. F., Wang, P., & Gonzalez, M. C. (2014). Development of origin-destination matrices using mobile phone call data. *Transportation Research Part C: Emerging Technologies*, *40*, 63–74. http://dx.doi.org/10.1016/j.trc.2014.01.002.

Itoh, S., & Hato, E. (2013). *Combined estimation of activity generation models incorporating unobserved small trips using probe person data10. (pp. 525–537), 525–537.*

Järv, O., Ahas, R., Saluveer, E., Derudder, B., & 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., & 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*, 122–135. http://dx.doi.org/10.1016/j.trc.2013.11.003.

Jiang, B., & Thill, J. -C. (2015). Volunteered geographic information: Towards the establishment of a new paradigm. *Computers, Environment and Urban Systems*, *53*, 1–3. http://dx.doi.org/10.1016/j.compenvurbsys.2015.09.011.

Joo, S., Oh, C., Jeong, E., & Lee, G. (2015). Categorizing bicycling environments using GPS-based public bicycle speed data. *Transportation Research Part C: Emerging Technologies*, *56*, 239–250. http://dx.doi.org/10.1016/j.trc.2015.04.012.

Kang, C., Gao, S., Lin, X., Xiao, Y., Yuan, Y., Liu, Y., & Ma, X. (2010). Analyzing and geo-visualizing individual human mobility patterns using mobile call records. *2010 18th International Conference on Geoinformatics* (pp. 1–7). http://dx.doi.org/10.1109/GEOINFORMATICS.2010.5567857.

Leuven (2016). Leuven in cijfers - Officiële site van de stad Leuven. Retrieved May 23, 2016. from http://www.leuven.be/bestuur/leuven-in-cijfers/.

Liu, X., Song, Y., Wu, K., Wang, J., Li, D., & Long, Y. (2015). Understanding urban China with open data. *Cities*, *47*, 53–61.

Liu, Y., Liu, X., Gao, S., Gong, L., Kang, C., Zhi, Y., & Chi, G. (2015). Social sensing: A new approach to understanding our socioeconomic environments. *Annals of the Association of American Geographers*, *105*(3), 512–530. http://dx.doi.org/10.1080/00045608.2015.1018773.

Magnanti, T. L., & Wong, R. T. (1984). Network design and transportation planning models and algorithms. *Transportation Science* (July 2017).

Manzoni, V., Maniloff, D., & Kloeckl, K. (2010). *Transportation mode identification and real-time CO_2 emission estimation using smartphones. Cambridge.*

Mc Fadden, D. L. (1978). Quantitative methods for analyzing travel behaviour of individuals: Some recent developments. In D. Stopher, & H. P. (Eds.), *Behavioural travel modelling* (pp. 279–318). London: Croom Helm London.

McInerney, J., Stein, S., Rogers, A., & Jennings, N. R. (2013). Breaking the habit: Measuring and predicting departures from routine in individual human mobility. *Pervasive and Mobile Computing*, *9*(6), 808–822. http://dx.doi.org/10.1016/j.pmcj.2013.07.016.

Montini, L., Prost, S., Schrammel, J., Rieser-Schüssler, N., & Axhausen, K. W. (2015). Comparison of travel diaries generated from smartphone data and dedicated GPS devices. *Transportation Research Procedia*, *11*, 227–241. http://dx.doi.org/10.1016/j.trpro.2015.12.020.

Nabais, J. L., Negenborn, R. R., Carmona Benítez, R. B., & Ayala Botto, M. (2015). Achieving transport modal split targets at intermodal freight hubs using a model predictive approach. *Transportation Research Part C: Emerging Technologies*, *60*, 278–297. http://dx.doi.org/10.1016/j.trc.2015.09.001.

Neis, P., & Zielstra, D. (2014). Recent developments and future trends in volunteered geographic information research: The case of OpenStreetMap. *Future Internet*, *6*(1), 76–106.

Ng, A. (2016). Stanford university. *Support vector machines - lecture notes* Retrieved November 17, 2016, from http://cs229.stanford.edu/materials.html.

Nitsche, P., Widhalm, P., Breuss, S., Brändle, N., & Maurer, P. (2014). Supporting large-scale travel surveys with smartphones – A practical approach. *Transportation Research Part C: Emerging Technologies*, *43*, 212–221. http://dx.doi.org/10.1016/j.trc.2013.11.005.

Nitsche, P., Widhalm, P., Breuss, S., & Maurer, P. (2012). A strategy on how to utilize smartphones for automatically reconstructing trips in travel surveys. *Procedia - Social and Behavioral Sciences*, *48*, 1033–1046. http://dx.doi.org/10.1016/j.sbspro.2012.06.1080.

Noulas, A. (2013). *Human urban mobility in location-based social networks: Analysis, models and applications.*

Pan, Y., Tian, Y., Liu, X., Gu, D., & Hua, G. (2016). Urban big data and the development of City intelligence. *Engineering*, *2*(2), 171–178.

Provost, F., & Fawcett, T. (2013). Data science and its relationship to big data and data-driven decision making. *Data Science and Big Data*, *1*(1), 51–59. http://dx.doi.org/10.1089/big.2013.1508.

Raper, J., Gartner, G., Karimi, H., & Rizos, C. (2007). A critical evaluation of location based services and their potential. *Journal of Location Based Services*, *1*(1), 5–45. http://dx.doi.org/10.1080/17489720701584069.

Rasmussen, T. K., Ingvardson, J. B., Halldórsdóttir, K., & Nielsen, O. A. (2015). Improved methods to deduct trip legs and mode from travel surveys using wearable GPS devices: A case study from the greater Copenhagen area. *Computers, Environment and Urban Systems*, *54*, 301–313. http://dx.doi.org/10.1016/j.compenvurbsys.2015.04.001.

Reddy, S., Burke, J., Estrin, D., Hansen, M., & Srivastava, M. (2008). *Determining transportation mode on mobile phones, 25–28.*

Reddy, S., Mun, M., Burke, J., Estrin, D., Hansen, M., & Srivastava, M. (2010). Using mobile phones to determine transportation modes. *ACM Transactions on Sensor Networks*, *6*(2), 1–27. http://dx.doi.org/10.1145/1689239.1689243.

Saelens, B. E., Sallis, J. F., & Frank, L. D. (2003). Environmental correlates of walking and cycling: Findings from the transportation, urban design, and planning literatures. *Annals of Behavioral Medicine*, *25*(2), 80–91. http://dx.doi.org/10.1207/S15324796ABM2502_03.

Safi, H., Assemi, B., Mesbah, M., & Ferreira, L. (2016). Trip detection with smartphone-assisted collection of travel data. *Transportation Research Record: Journal of the Transportation Research Board*, *2594*, 18–26. http://dx.doi.org/10.3141/2594-03.

Schlkopf, B., & Smola, A. J. (2001). *Learning with kernels: Support vector machines, regularization, optimization, and beyond (adaptive computation and machine learning).* Cambridge: The MIT Press.

Seidl, D. E., Jankowski, P., & Tsou, M. -H. (2016). Privacy and spatial pattern preservation in masked GPS trajectory data. *International Journal of Geographical Information Science*, *30*(4), 785–800. http://dx.doi.org/10.1080/13658816.2015.1101767.

Semanjski, I. (2015). Potential of big data in forecasting travel times | Potencijal velikih setova podataka u prognoziranju vremena putovanja. *Promet - Traffic - Traffico*, *27*(6).

Semanjski, I., Bellens, R., Gautama, S., & Witlox, F. (2016). Integrating big data into a sustainable mobility policy 2.0 planning support system. *Sustainability (Switzerland)*, *8*(11), 1–19. http://dx.doi.org/10.3390/su8111142.

Semanjski, I., & Gautama, S. (2015). Smart City mobility application—Gradient boosting trees for mobility prediction and analysis based on Crowdsourced data. *Sensors*, *15*(7), 15974–15987. http://dx.doi.org/10.3390/s150715974.

Semanjski, I., & Gautama, S. (2016a). *Crowdsourcing mobility insights – Reflection of attitude based segments on high resolution mobility behaviour data.* Transportation Research Part C: Emerging Technologies, 71. http://dx.doi.org/10.1016/j.trc.2016.08.016.

Semanjski, I., & Gautama, S. (2016b). Sensing human activity for smart cities' mobility management. In I. N. Da Silva, & R. A. Flauzino (Eds.), *Smart cities technologies* (pp. 1–26). Rijeka, Croatia: InTech. http://dx.doi.org/10.5772/65252.

Shin, D., Aliaga, D., Tunçer, B., Arisona, S. M., Kim, S., Zünd, D., & Schmitt, G. (2015). Urban sensing: Using smartphones for transportation mode classification. *Computers, Environment and Urban Systems*, *53*, 76–86. http://dx.doi.org/10.1016/j.compenvurbsys.2014.07.011.

Steininger, K., Vogl, C., & Zettl, R. (1996). *Car-sharing organizations in mobility behavior3(4). (pp. 177–185), 177–185.*

Stenneth, L., Wolfson, O., Yu, P. S., Xu, B., & Morgan, S. (2011). *Transportation mode detection using mobile phones and GIS information.*

Stopher, P. R., & Greaves, S. P. (2007). Household travel surveys: Where are we going? *Transportation Research Part A: Policy and Practice*, *41*(5), 367–381. http://dx.doi.org/10.1016/j.tra.2006.09.005.

Stopher, P. R., & Wilmot, C. G. (2000). Some new approaches to designing household travel surveys-time-use diaries and GPS. *Paper presented at the 79th Annual TRB meeting. Washington DC.*

Tabuchi, T. (1993). Bottleneck congestion and modal split. *Journal of Urban Economics*, *34*(3), 414–431.

Thorsby, J., Stowers, G., Wolslegel, K., & Tumbuan, E. (2017). Understanding the content and features of open data portals in American cities. *Government Information Quarterly*, *34*(1), 53–61.

Toole, J. L., Colak, S., Sturt, B., Alexander, L. P., Evsukoff, A., & González, M. C. (2015). The path most traveled: Travel demand estimation using big data resources. *Transportation Research Part C: Emerging Technologies*, *58*, 162–177. http://dx.doi.org/10.1016/j.trc.2015.04.022.

Toole, J. L., Ulm, M., González, M. C., & Bauer, D. (2012). Inferring land use from mobile phone activity. *Proceedings of the ACM SIGKDD International Workshop on Urban Computing - UrbComp '12*, 1. http://dx.doi.org/10.1145/2346496.2346498.

Turner, S., Eisele, W., Benz, R., & Holdener, D. (1998). *Travel time data collection handbook.* Arlington: Texas Transportation Institute.

Vapnik, V. (1998). *Statistical learning theory.* New York: Wiley.

Vapnik, V., & Yakovlevich, A. (1963). Pattern recognition using generalized portrait method. *Automation and Remote Control*, *24*, 774–780.

Vij, A., & Shankari, K. (2015). When is big data big enough? Implications of using GPS-based surveys for travel demand analysis. *Transportation Research Part C: Emerging Technologies*, *56*, 446–462. http://dx.doi.org/10.1016/j.trc.2015.04.025.

Vlahogianni, E. I. (2015). Optimization of traffic forecasting: Intelligent surrogate modeling. *Transportation Research Part C: Emerging Technologies*, *55*, 14–23. http://dx.doi.org/10.1016/j.trc.2015.03.016.

Vlahogianni, E. I., Park, B. B., & Van Lint, J. W. C. (2015). Big data in transportation and traffic engineering. *Transportation Research Part C: Emerging Technologies*, *58*, 161. http://dx.doi.org/10.1016/j.trc.2015.08.006.

Wang, H., Calabrese, F., Lorenzo, G. D., & Ratti, C. (2010). *Transportation mode inference from anonymized and aggregated mobile phone call detail records, 318–323.*

Wang, J., & Shi, Q. (2013). Short-term traffic speed forecasting hybrid model based on chaos – Wavelet analysis-support vector machine theory. *Transportation Research Part C*, *27*, 219–232. http://dx.doi.org/10.1016/j.trc.2012.08.004.

Wang, S., Chen, C., & Ma, J. (2010). Accelerometer based transportation mode recognition on mobile phones. *Asia-Pacific Conference on Wearable Computing Systems*, *2010*, 44–46. http://dx.doi.org/10.1109/APWCS.2010.18.

Wang, S., Chen, Y., & Chen, Z. (2013). Recognizing transportation mode on mobile phone using probability fusion of extreme learning machines. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *21*(supp02), 13–22. http://dx.doi.org/10.1142/S0218488513400126.

Williams, N. E., Thomas, T. A., Dunbar, M., Eagle, N., & Dobra, A. (2015). Measures of human mobility using mobile phone records enhanced with GIS data. *PloS One*, *10*(7), 1–16. http://dx.doi.org/10.1371/journal.pone.0133630.

Witlox, F. (2007). Evaluating the reliability of reported distance data in urban travel behaviour analysis. *Journal of Transport Geography*, *15*(3), 172–183. http://dx.doi.org/10.1016/j.jtrangeo.2006.02.012.

Witlox, F. (2015). Beyond the data Smog ? *Transport Reviews*, *35*(3), 245–249. http://dx.doi.org/10.1080/01441647.2015.1036505.

Wolf, J., Bricka, S., Ashby, T., & Gorugantua, C. (2004). Advances in the application of GPS to household travel surveys. *Household travel survey* Retrieved from http://onlinepubs.trb.org/onlinepubs/archive/conferences/nhts/Wolf.pdf.

Wolf, J., Guensler, R., & Bachman, W. (2001). Elimination of the travel diary: Experiment to derive trip purpose from global positioning system travel data. *Transportation Research Record*, *1768*(1), 125–134. http://dx.doi.org/10.3141/1768-15.

Wu, Y., Lee, Y., & Yang, J. (2008). Robust and efficient multiclass SVM models for phrase pattern recognition. *Pattern Recognition*, *41*(2008), 2874–2889. http://dx.doi.org/10.1016/j.patcog.2008.02.010.

Xia, H., Qiao, Y., Jian, J., & Chang, Y. (2014). Using smart phone sensors to detect transportation modes. *Sensors (Basel, Switzerland)*, *14*(11), 20843–20865. http://dx.doi.org/10.3390/s141120843.

Xiao, G., Juan, Z., & Zhang, C. (2015). Travel mode detection based on GPS track data and Bayesian networks. *Computers, Environment and Urban Systems*, *54*, 14–22. http://dx.doi.org/10.1016/j.compenvurbsys.2015.05.005.

Xiao, Z., Wang, Y., Fu, K., & Wu, F. (2017). Identifying different transportation modes from trajectory data using tree-based ensemble classifiers. *ISPRS International Journal of Geo-Information*, *6*(3), 57. http://dx.doi.org/10.3390/ijgi6020057.

Xu, Y., Shaw, S. -L., Zhao, Z., Yin, L., Lu, F., Chen, J., ... Li, Q. (2015). Understanding aggregate human mobility patterns using passive mobile phone location data- a home-based approach understanding aggregate human mobility patterns using passive. *Transportation July*, *106*(2), 489–502. http://dx.doi.org/10.1007/s11116-015-9597-y.

Xu, Y., Shaw, S., Zhao, Z., Yin, L., Lu, F., Chen, J., ... Fang, Z. (2016). Another tale of two Cities: Understanding human activity space using actively tracked cellphone location data. *Annals of the Association of American Geographers*, *106*(2), 489–502. http://dx.doi.org/10.1080/00045608.2015.1120147.

Zhao, P., Jia, T., Qin, K., Shana, J., & Jiao, C. (2015). Statistical analysis on the evolution of OpenStreetMap road networks in Beijing. *Physica A*, *420*, 59–72.

Zhou, X., Yu, W., & Sullivan, W. C. (2016). Making pervasive sensing possible: Effective travel mode sensing based on smartphones. *Computers, Environment and Urban Systems*, *58*, 52–59. http://dx.doi.org/10.1016/j.compenvurbsys.2016.03.001.