CISMOB
Interreg Europe

universidade de aveiro

Technical paper 6

Exploring crowdsourcing information to predict traffic-related impacts

Authors P. Tafidis , J. Teixeira, B. Bahmankhah, E. Macedo, M.C. Coelho, J. Bandeira

PP1

1 Introduction

The increasing road transport volumes in urban areas are a primary source of air pollution and greenhouse gas emissions [1]. The conditions are usually worst in areas that generate and attract many trips, such as city centres, shopping areas etc. In some cases, the low level of public transport services or even more the absence of alternative transport modes encourage the use of private vehicles aggravating the situation even more.

Advance traffic management systems are essential tools to reduce air emissions and promote sustainable mobility. They integrate multiple technologies to improve traffic flow and reduce congestion. High quality real-time traffic data is necessary to maintain the effectiveness and reliability of traffic management services. In transport systems, traditional data collection systems is usually considered costly and lengthy and limited to specific areas of interest.

The aim of this paper is to explore an alternative source of data by examining if there is any correlation between the information provided by applications like Google Maps regarding the popular times of specific areas and the amount of air emissions produced during that period.

European Union
European Regional
Development Fund

2 Literature Review

Many initiatives tried to estimate traffic and environmental impacts using new sources of information. For a case study in Chicago, a methodology was developed based on Bing Maps information to estimate traffic [2]. In [3], traffic flow predictions were based on the use of Twitter data. A model was developed using features from tweets to estimate features social media information have used to predict traffic flow using tweet features. A model has been developed, based on tweets features in order to estimate traffic flow in short terms before sport game events.

Reference [4] proposed a probabilistic data analysis model, based on time and space, to specify habitual and non-habitual overcrowding hotspots in public transportation systems. The model could be able to predict crowd hotspots under assumption data from web-sources information. The authors used related statistics data from smartcards that are used in Singapore's public transport system and from social networks as well. The results showed that the extracted data are intuitively plausible and consistent both locally and globally.

In [5], authors suggested a methodology with the use of social media source data to estimate the Origin-Destination travel demand, while in [6], authors have used four different social media source information in order to create a model for travel demand. Although these data from social media cannot be used to create accurate demand model, they can be used as complementary data sets to enrich the conventional transport-related collection methods.

In this context, further information have provided in [7] regarding the potential role of popular online web sources (Bing map, Google map, Baidu map, OpenStreetMap, etc.) and social media data on traffic information extraction and visualization, traffic event detection, traffic information prediction, and traffic sentiment analysis. The authors concluded that the effect of social networks on transportation area will be significant in future.

In the last few years, plenty of new applications are using crowdsourcing to give users real-time information through graphs or photos about the traffic of specific locations. For instance, Google launched in July 2015 a new feature in Google Maps that allows users to know a certain place's busy times by crowdsourcing users' visits. The new update aims at helping people in their decision-making process about when it is the best time to visit a business, a restaurant, a shop etc. The data usually includes: i) popular times graph per day and hour showing how busy is a specific location based on average popularity over the last several weeks, ii) live activity data updating by real-time information and easily comparable with the average values and, iii) visit duration, showing the average spending time of people at the place. This sort of information could be especially useful as an alternative source of predicting traffic related impacts in cities or areas where floating car data (FCD) penetration rates are insufficient or traffic-monitoring stations are limited.

3 Methodology

3.1 Case Studies

In this paper, six road links are examined near two important commercial areas in the city of Aveiro, Portugal. Links L1 and L2 consist of the main entrance and exit to The Aveiro Shopping Center and links L3 and L4 connect the city of Aveiro with the industrial zone (Fig. 1). The area is characterized by industrial land use, the studied links are between roundabouts and there are various unsignalized intersections with minor roads. Regarding the second study area, link L5 is the main street that lead to Centro Comercial Glicínias Plaza, while link L6 leads out of it (Fig. 2). The land use of the area is mainly residential and the studied links are between two-lane roundabouts. One crosswalk interrupts them, while in link L6 there is the only entrance and exit to a gas station. To enhance the applicability of the proposed methodology, we chose links that clearly serve as traffic distributors to the commercial zones but also, due to their proximity to important arterials, the traffic performance can be affected with traffic from different Origin-Destination pairs.

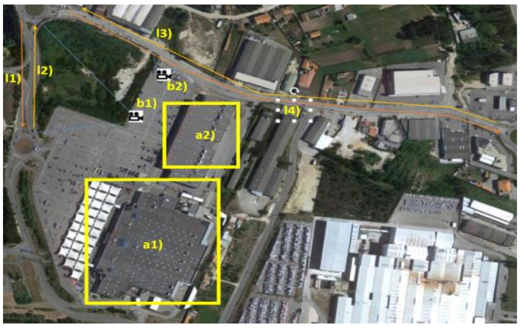

Fig. 1. Study area of Aveiro Shopping Center

Fig. 2. Study area of Centro Comercial Glicínias Plaza

3.2 Data Collection

For the purpose of the paper three different data sets were collected:

- Traffic volumes in 15 minutes intervals with the use of cameras.
- Traffic dynamics (travel time, speed, and acceleration) with the use of a light-duty vehicle equipped with a GNSS data logger to collect second by second trajectory data required for microscopic analysis.
- Crowdsourcing information in real time (Popular Times) from Google Maps regarding the activity of the commercial areas (Fig. 3). Google Maps present Popular Times with the use of a bar chart without providing values, For the purpose of this paper, we assume that the minimum value of the bar is zero and the maximum is one and we divided it in ten equal parts giving them the respective values.

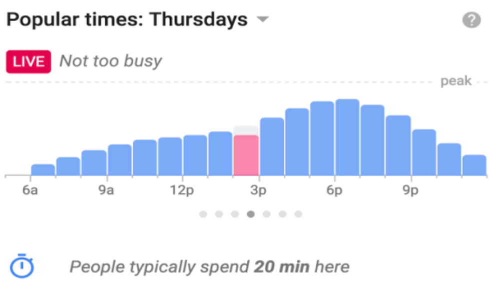

Fig. 3. Example of Popular Times feature of Google Maps

The data collection period was chosen in order to obtain a diversified range of demand in both case studies, including normal weekday and special weekend conditions and lasted 6 hours each day. Regarding the vehicle dynamic monitoring, 10 runs were performed per hour for each link by different drivers to increase the heterogeneity of driving behavior [8]. The probe vehicle was moving according to the driver's perception of the average speed of the traffic stream [9]. A MATLAB routine developed by one of the authors to organize data and calculate dynamic data obtained during field tests and write them in Excel was used. Excel was used to perform data exploratory analysis, mainly focused in stablishing associations between variables.

3.3. Emission Estimation

The Emissions estimation was based on the concept of Vehicle Specific Power (VSP). The VSP model reflects the comparison of driving behavioral effects in fuel consumption and vehicle emissions for different air pollutants:

- Carbon Dioxide (CO_2);
- Carbon Monoxide (CO);
- Nitrogen Oxides (NO_x);
- Hydrocarbon (HC).

The VSP represents the power required to the engine based on the road gradient, aerodynamics, kinetic energy and friction to the movement, and is a model that has proven to be very effective in estimating emissions from petrol and diesel cars [8]. The VSP mode can be expressed as

$$VSP = v[1.1a+9.81(atan(sin\ (grade)))+0.123]+0.000302v^3 \quad (1)$$

where: v = vehicle speed (m/s), a = vehicle acceleration/deceleration rate (m/s^2), grade = vehicle vertical rise divided by the horizontal run (%). Each VSP bin refers to one of 14 modes. Each VSP mode is defined by a range of VSP values which are associated to an emission rate. Each calculation of VSP results in a unique classification to a VSP mode [10,11]. The following distribution fleet composition was considered based on the Portuguese vehicle classification for the case study: 38% of light duty gasoline vehicles and 62% of light duty diesel vehicles [12]. It should be noted that due to the flat terrain, the grade of road was considered negligible. Due to space limitations, we will focus our analysis on CO_2 and NO_X emissions, a greenhouse gas and a local pollutant critical in the study region.

4 Results

Table I displays the main characteristics of each link after the analysis of the results. Fig. 4 presents the two approaches that were followed to find the relationships between the traffic volumes and the Popular Times of the commercial zones. The standard interval of 15 minutes was used in this study as it is the lowest interval which flow rates are statistical stable [13].

In the first approach, traffic volumes of each interval were compared to popular times' value in the end of the respective time period, while in the second the comparison was made with the value in the end of the next 15 minutes period. Although the tests using the second approach did not provided significant improvement on the results obtained for the first, we only report the first approach. The main objective of this study is to explore the potential of crowdsourcing information to contribute as an alternative source of data for real-time traffic estimation and traffic-related impacts.

TABLE I. CHARACTERISTICS OF LINKS

Links	Length (m)	Q_{max} (volume per hour)	Average Q (volume per hour)
L1	215	816	544
L2	210	924	608
L3	718	210	210
L4	722	210	210
L5	88	700	700
L6	90	700	700

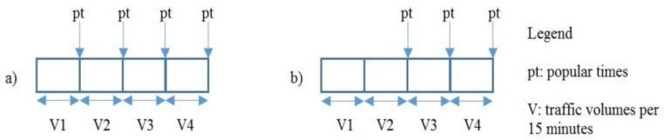

Fig. 4. First (a) and second (b) approach for correlation analysis

4.1. Centro Comercial Glicínias Plaza

1) Relationships between various variables and Popular Times

To study the relationships between Traffic Volumes, Travel time, Emissions and Popular Times, we try to use simple models, such as linear and quadratic fittings, in order to ease the interpretation of results. We are interested in whether Popular Times, as a predictor variable, can explain the variability of other variables. From Table II, the first observation is that the results for the first day are not so encouraging, while the second day presented better significance values (p-value). This means that the chosen models are adequate and variables are correlated with Popular Times, i.e., changes in Popular Times (predictor) are associated with changes in the other variables (response).

Considering both days results, we can see that Traffic Volumes and CO_2 emissions are the variables that Popular Times can explain better, since they present highest coefficient of determination values. Better results are reported in second day for Link 6, where we can see that Traffic Volumes is highly linearly correlated with Popular Times, concretely, 90% of the variability of Volumes can be explained by Popular Times. We can also find that more than 61% of the variation in CO_2 emissions, and more than 51% in Travel time and NO_x emissions, can be explained by Popular Times. These results support the idea that ICT applications can be used to predict traffic-related impacts.

TABLE II. MODEL PARAMETERS BETWEEN RESPONSE VARIABLES AND POPULAR TIME

Links	**Variables**	**Coefficients of the model**			**R^{2a}**	**p-value**
		a	***b***	***c***		
1st day						
L5	Volumes	2622.9	-2059	908.75	0.21	3.78E-04
	Travel Time	4504.6	6502.3	linear fitting	0.04	1.21E-01
	System CO_2	6744.9	11556	linear fitting	0.05	6.88E-02
	System NO_x	15.016	25.805	linear fitting	0.05	7.42E-02
L6	Volumes	342.32	417.73	linear fitting	0.44	2.17E-10
	Travel Time	162709	-121358	36181	0.12	1.31E-02
	System CO_2	139586	-99243	39538	0.13	7.63E-03
	System NO_x	374.81	-270.05	108.48	0.12	1.27E-02
2nd day						
L5	Volumes	4933.6	-3372.2	881.91	0.44	3.14E-10
	Travel Time	77890	-50449	12499	0.41	2.70E-09
	System CO_2	76797	-42000	12741	0.41	2.50E-09
	System NO_x	120.92	-51.293	20.812	0.34	1.71E-07
L6	Volumes	1372.7	-177.29	linear fitting	0.91	7.78E-42
	Travel Time	37315	-6703.2	linear fitting	0.52	5.63E-14
	System CO_2	55962	-10426	linear fitting	0.61	1.34E-17
	System NO_x	158.1	-29.962	linear fitting	0.52	1.87E-13

[a.] Coefficient of Determination

2) **Correlations between Traffic Volumes and Popular Times**

Fig. 5 presents scatterplots highlighting the relationships between Traffic Volumes per hour (vph) near Centro Comercial Glicínias Plaza and Popular Times for the two days.

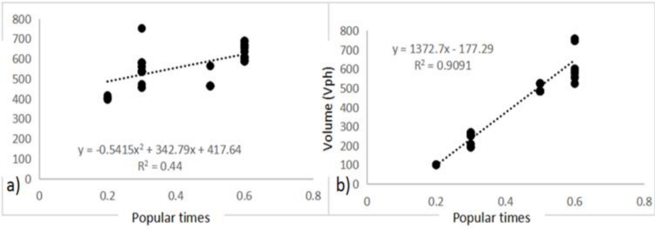

Fig. 5. a) Linear correlation for link L6 in first day b) Linear correlation for link L6 in second day

For the sake of clarity, we are aware that no accurate assertion can be made in terms of clear correlations between such variables when values are registered at the same time. Evidently, one cannot say that in the moment we registered a vehicle, the people on it are already inside the shopping. Nevertheless, we believe that there is, however, some relationship between the traffic volume and the volume of people inside the shopping center within a delay of few minutes. In the first day of study, correlations presented lower values. This was due to a considerable affluence to the gas station nearby, leading to higher traffic volumes during low values of popular hours (0.2 and 0.3). In this light, better results were obtained during the weekend than the weekday. Concretely, in the second day, there can be verified a linear fitting, where 90% of Traffic Volume can be explained by Popular Times, while regarding the first day, with a linear fitting we are only able to explain 44% of the variability. Obviously, we could improve the adjustment in day 1 by using a third-order polynomial fitting. From the results, we can assume that higher percentage of vehicles during the second day had as origin or destination the commercial area.

3) **Correlations between Emissions and Popular Times**

Fig. 6 shows the correlation between System CO_2 emissions (where it considers the Portuguese vehicle composition fleet and the traffic volume registered in each day) and System NOx emissions with Popular Times for each day.

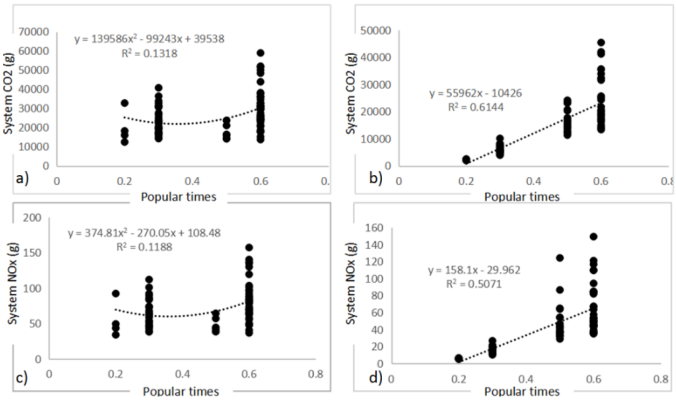

Fig. 6. a) c) Quadratic correlation for NOx and CO_2 for first day in Link L6; b) d) Linear correlation for NOx and CO_2 for second day on Link L6

There is no linear correlation on the first day. In this case, the emission values are too high with respect to Popular Times values, which are between 0.2 and 0.3 where traffic is not supposed to be so high. Considering the results on de second day, we can observe that there is higher correlation between emission values and Popular Times. This means that under similar circumstances Popular Times can be a valid explanatory variable to predict road traffic emissions.

4.2. Aveiro Shopping Center

1) **Relationships between various variables and Popular Times**

In Table III, the first observation is that we have better results, not only in terms of the coefficient of determination value, but also in terms of significance, when compared to Glicínias. In particular, we were able to estimate linear fittings for almost all cases.

TABLE III. COEFFICIENT OF DETERMINATION BETWEEN EMISSIONS AND TRAVEL TIME VARIABLES

Links	Variables	Coefficients of the model			R^2	p-value
		a	*b*	*c*		
L1	Volumes	942.31	85.737	linear fitting	0.67	8.74E-16
	Travel Time	16284	-142.68	linear fitting	0.5	1.61E-12
	System CO_2	27173	2950.1	linear fitting	0.67	6.41E-16
	System NO_x	75.802	10.962	linear fitting	0.38	1.14E-07
L2	Volumes	1097	74.875	linear fitting	0.80	8.47E-23
	Travel Time	18914	315.54	linear fitting	0.55	3.99E-12
	System CO_2	33287	4141	linear fitting	0.62	4.07E-14
	System NO_x	-185	280.97	-26.94	0.42	1.05E-07
L3	Volumes	1005.6	-73.07	linear fitting	0.71	4.69E-17
	Travel Time	39230	-1.9003	4019.2	0.56	7.59E-11
	System CO_2	137570	-15920	linear fitting	0.59	8.22E-13
	System NO_x	341.72	-30.624	linear fitting	0.44	8.06E-09
L4	Volumes	1051.4	-13.138	linear fitting	0.57	1.54E-12
	Travel Time	75585	-7482.7	linear fitting	0.52	7.57E-11
	System CO_2	142103	-10772	linear fitting	0.60	1.98E-13
	System NO_x	393.23	-23.937	linear fitting	0.55	7.82E-12

We can see that for links 1, 2 and 3, the estimation of the variability of NO_X emissions by Popular Time is smaller than that obtained for the other response variables, nevertheless, we can assert that more than 38% of NO_X emissions can be explained by Popular Times. Considering all the links, we can see that Traffic Volume and CO_2 emissions are the response variables that Popular Times can explain better. In particular, Popular Times can explain approximately 67% of Traffic Volumes and CO_2 emissions in Link 1. Concerning Link 2, 80% and almost 62% of the variability of Traffic Volumes and CO_2 emissions, respectively, can be explained by Popular Times, while in Link 3, such values are closer to 71% and 59%, respectively. For Link 4, we see that more than 50% of the variability of all response variables can be explained by Popular Times. Once again, the results show that we can estimate traffic-related impacts by using ICT applications.

It can be pointed out that for Popular Times values between 0.2 and 0.6, it was drivers behavior that most influenced the level of emissions, while for values between 0.7 to 0.8, it was the Travel time. It should also kept in mind that when there are registered low levels of Popular Times values, it could mean that more people is leaving the shopping area, thus, increasing the existing traffic and justifying the relation with drivers behavior. As well as in the case of higher Popular Times values, the weight of travel time is explained on the fact that more people are coming in than going out, meaning less vehicles exiting the commercial area.

2) **Correlations between Traffic Volumes and Popular Times**

Fig. 7 shows the correlations between Traffic Volumes and Popular Times values for the Aveiro Shopping Center. It is important to notice that the Popular Times value 0.5 there was a high variation of traffic volumes in all links. Link L3 and L4 offer one lane per direction, while link L1 and L2 offer two lanes per direction. In case of L1 and L2, that represent the main entrance to the commercial area, it was expected a high correlation.

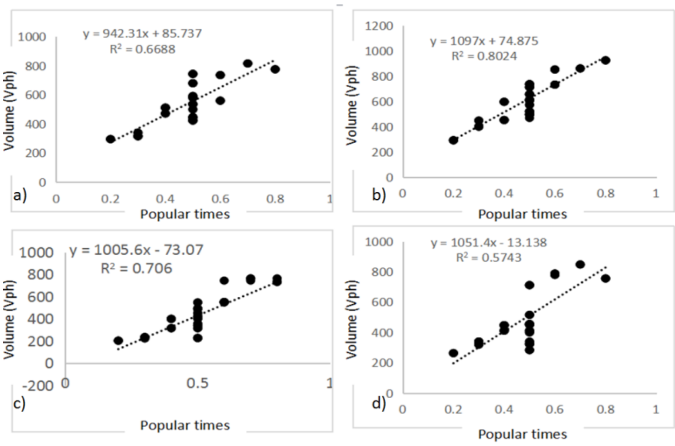

Fig. 7. Correlations between Traffic Volumes and Popular Times a) Link L1; b) Link L2; c) Link L3; d) Link L4

3) **Correlations between Emissions, Travel time and Popular Times**

Fig. 8 displays low correlation between emissions and Travel time with Popular Times compared with the case of Centro Comercial Glicínias Plaza in second day. These results can be explained by the drivers behaviour, which was shown to be inconsistent and aggressive in case of L1, thus producing more emissions, meaning that without traffic congestion, drivers were able to experience greater accelerations and speeds, such as greater speed reductions near the roundabouts. With some traffic, the drivers changed their behavior and begun to adopt a median driving style. We can also refer that the higher slope of the fitting, for low popular hour, is concerning the emissions. Popular Times valued 0.5, where the drivers changed their driving style, is a key point with major impact on the correlations. If we neglect this point, correlations for CO_2 go up to 0.79 and NOx to 0.67.

In L1 driver's behavior was the most important factor (in both cases, with or without traffic). Link L3 gave correlations lower than with L1. This is essentially because of Popular Times valued 0.8 that also appears in the peak hours at late afternoon, where many cars in L3 were mostly entering or exiting from Aveiro (once again it is important to refer that in this link driver's behaviour is strongly related to traffic volume). If Popular Times valued 0.8 is taken out of the correlations, then CO_2 correlation values will be 0.71, NOx will be 0.50.

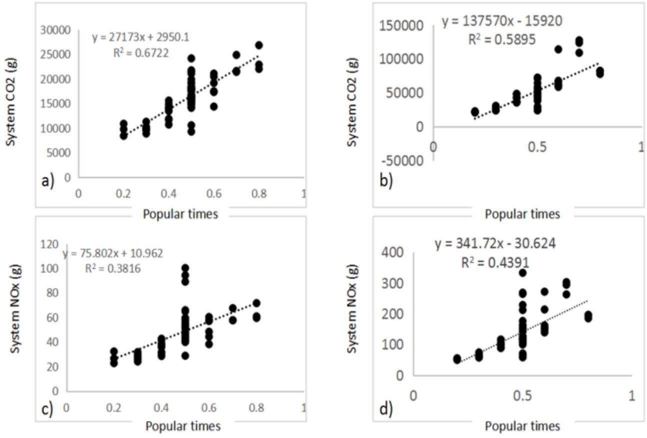

Fig. 8. a,c) Correlations between emissions and Popular Times of link L1; b,d) Correlations between emissions and Popular Times of link L3

Conclusions

The paper explores traffic and environmental impacts using ICT applications, namely Google Maps Popular Times. For this purpose, we focus this study on two shopping areas in Aveiro. Certain values given by Google have less tests made than others, implying that the conclusions or observations are dependent of the low population studied in those cases.

The analysis of the statistical fitting of data suggests that for all the tested links we can establish clear relationships (most of them linear) between Traffic Volumes, Travel time, Emissions, and Popular Times. Regarding Aveiro Shopping Center, correlations between Volume and Popular Time were above 0.65, except in L4. In the shortest links L1 and L2 we obtained better results than in L3 and L4, because at peak hours the majority of road users use those links to get out of the city and not go to the commercial area. Higher correlations were obtained during weekend when a higher percentage of traffic has as destination the analysed recreation in shopping areas.

With this in mind, it was proven the relation and potential use of crowdsourcing information to describe traffic performance and emissions in nearby links. The preliminary results are promising, however there is indeed a need for a better calibration between the time data is collected and the time Google supplies the information. Although the potential of crowdsourcing information, we should mention that this is not prepared to relate accurately atypical data without predicting, studying and working some kind of adaptive learning algorithm to identify and eliminate/predict similar cases (where the usage of the link by the road users is used to attain other destiny than the place yielding the information).

The presented results support our goal, showing that ICT technologies/applications can be used to estimate environmental impacts. Further research is going to be conducted in different areas, by considering integrated crowdsourcing data of multiple sites (not only recreational and shopping areas) and regions with heterogenic characteristics.

Acknowledgements

The authors acknowledge the support of Project CISMOB (PGI01611), funded by the Interreg Europe Programme, the Strategic Project UID-EMS-00481-2013 and @CRUiSE project (PTDC/EMS-TRA/0383/2014), funded within the Project 9471 – Reforçar a Investigação, o Desenvolvimento Tecnológico e a Inovação (Project 9471 – RIDTI) and supported by the European Community Fund FEDER. J. Bandeira also acknowledges the support of FCT for the Scholarship SFRH/BPD/100703/2014.

References

[1] R. Colvile, E. Hutchinson, J. Mindell and R. Warren, “The transport sector as a source of air pollution”, in Atmospheric Environment, vol. 35(9), pp. 1537-1565, March 2001.

[2] A.I.J. Tostes, F. de LP Duarte-Figueiredo, R. Assunção, J. Salles and A.A. Loureiro, “From data to knowledge: city-wide traffic flows analysis and prediction using bing maps,” in the Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing, pp. 12, August 2013.

[3] M. Ni, Q. He and J. Gao, “Using social media to predict traffic flow under special event conditions,” 2014 [Transportation Research Board 93rd Annual Meeting].

[4] F.C. Pereira, F. Rodrigues, E. Polisciuc and M. Ben-Akiva, M., “Why so many people? explaining nonhabitual transport overcrowding with internet data” IEEE Transactions on Intelligent Transportation Systems, vol. 16(3), pp.1370-1379, 2015.

[5] F. Yang, P.J. Jin, X. Wan, R. Li and B. Ran, “Dynamic origin-destination travel demand estimation using location based social networking data,” 2014 [Transportation Research Board 93rd Annual Meeting].

[6] E. Chaniotakis, C. Antoniou, J.M.S. Grau and L. Dimitriou, “Can Social Media data augment travel demand survey data?,” [IEEE 19th International Conference on Intelligent Transportation Systems pp. 1642-1647, November 2016].

[7] Y. Lv, Y. Chen, X. Zhang, Y. Duan and N.L. Li, “Social media based transportation research: the state of the work and the networking,” in IEEE/CAA Journal of Automatica Sinica, vol. 4(1), pp.19-26, 2017.

[8] R. Downling, A. Skabardonis, V. Alexiadis, “Traffic Analysis Toolbox Volume III: Guidelines for Applying Traffic Microsimulation Modeling Software,” 2004.

[9] S.M. Turner, W.L. Eisele and R.J. Benz, “Travel time data collection handbook,” March 1998.

[10] M.C. Coelho, H.C. Frey, N.M. Rouphai, H. Zhai and L. Pelkmans, “Assessing methods for comaring emissions from gasoline and diesel light-duty vehicles based on microscale measurements,” in Transportation Research Part D: Transport and Environment, vol.14(2), pp. 91-99, 2009.

[11] H.C. Frey, A. Unal, J. Chen, S. Li and C. Xuan, “Methodology for developing modal emission rates for EPA’s multi-scale motor vehicle & equipment emission system,” Ann Arbor, Michigan: US Environmental Protection Agency, 2002.

[12] Emisia SA., COPERT data, Retrieved April 9, 2017, from http://emisia.com/products/copert-data.

[13] D. Florea, D. Covaciu, I. Preda and J. Timar, “Traffic counts methodologies for various type of road traffic applications”.