Traffic Flow Estimation for Urban Roads Based on Crowdsourced Data and Machine Learning Principles

Sakitha P. Kumarage(✉), R. P. G. K. S. Rajapaksha, Dimantha De Silva, and J. M. S. J. Bandara

Department of Civil Engineering, University of Moratuwa, Moratuwa, Sri Lanka
{ra-sakitha, l58002T, dimanthads, bandara}@uom.lk

Abstract. The congestion in urban road networks are common problem across all urban centers. Understanding the traffic flow across the road segments are necessary to provide viable solutions, but a very expensive task specially for developing countries. This paper proposes an economical approach for a directional flow prediction model for urban road based on Google Distance Matrix API data, archived traffic flow data, and geometric data. Data gathered was aggregated in space and time as attributes to the model estimation. Deviating from traditional probability estimation, a K- Nearest Neighbour regression method was used in the analysis. The model is validated using a test dataset which showed a root mean square error and a mean absolute error of prediction as 9.479 and 2.318, which suggest that with the use of travel time and speed data gathered from Google Distance matrix API is possible to estimate lane flow when road geometry is defined.

Keywords: Traffic flow estimation · Machine learning · KNN regression Google APIs · Crowdsourced data

1 Introduction

Traffic flow is a key contributor for traffic management and control systems [1]. Availability of traffic flow data for urban road networks will enable planning and decision making efficient for transport planners and policy makers [1]. However, the collection of consistent traffic flow data for an urban road network in developing countries has been challenged due to funding gaps [2]. In most of the developed countries, the flow data could be obtained from traffic sensors and surveillance systems embedded in road networks [2], which become an expensive alternative for developing countries to prioritize.

With the improvement of intelligent transportation systems, traffic flow prediction has become a major consideration [3]. Communication and detection methodologies have phased up with the advancement of crowdsourced data mining, allowing more readily extractable information on transport and mobility [3]. Further use of crowdsourced data has become more consistent and reliable as it enables to collect a large amount of data samples continuously [4]. The investment in infrastructure is minimized

T. Kováčiková et al. (Eds.): INTSYS 2017, LNICST 222, pp. 263–273, 2018.
https://doi.org/10.1007/978-3-319-93710-6_27

[3], resulting crowdsourced data as a low-cost alternative since the crowdsourced data collection techniques are based on consumer services such as telephone calls, GPS navigation, Geotagged data transfer etc. [4]. Google Distance Matrix API provides travel time for a given origin and destination based on the above concept [5].

The objective of this study is to develop a non-parametric non-linear traffic flow estimation model based on K- nearest neighbour regression, which uses travel time and speed data obtained from Google Distance Matrix API and road geometry data. The study includes a detailed literature survey, methodology identification, data validation and results of the analysis model.

2 Literature Review

2.1 Traffic Flow Estimation

Traffic Flow Rate (q) is defined as a number of vehicles passing a point on a space during a unit period of time, which is given by Eq. 1 [6].

$$q = \frac{n}{T} = \frac{n}{\sum_{i=1}^{n} h_i} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} h_i} = \frac{1}{\overline{h}} \tag{1}$$

q = flow/volume;
n = Number of vehicles;
T = time duration;
h = time head way;
$\overline{h}$ = mean time headway.

In 1936, Greenshield assumed a parabolic flow-density relation corresponding to a linear speed-density relation [6]. Improving from that, transport specialists developed different mathematical models for uninterrupted traffic flow considering macroscopic and microscopic traffic flow characteristics [6]. Microscopic traffic simulators [e.g., MITSIMLab, AIMSUN, VISSIM] involve detailed models of driver behavior, comprising car-following, gap-acceptance, lane-changing, and other disaggregate behavioral models [6]. In developing countries, due to the existing heterogeneous traffic nature and complexity in obtaining data, it is difficult to use microscopic simulations [6]. Vehicle distribution of developing countries shows a higher percentage of motorized two-wheelers and three-wheelers compared to four-wheelers [7]. Hence traffic flow characteristics are much different when compared with developed countries which have more than 80% motor cars in traffic flow [7]. Further, it could be observed that two-lane roads are abundant in developing countries [7]. In two-lane roads lane changing and passing maneuvers typically performed when sight distance and gaps being available in the opposing traffic stream [7]. Hence the directional flow is influenced by opposite flow which has to be taken into consideration when characterizing directional flow [7]. Chandra has found that road width, shoulder width, and directional split are significantly affected on the free flow speed and capacity of two-lane highways [7]. Moreover,

classical methods of traffic flow analysis omit the temporal variation of traffic flow [8]. Although incidental analyses are possible, analysing long-term behavior is challenging with classical method. Hence researchers looked at time series analysis and machine learning principles to incorporate spatiotemporal variables in traffic flow prediction [8].

2.2 Crowdsourced Data

Crowdsourced data is a novel approach in gathering information from smartphone users, social media and the internet in general [4]. As suggested by Chatzimilioudis et al. collection of crowd locations, mobility speed, travel frequency data become viable with the widespread availability and the multi-sensing features of smartphones such as geolocation, movement, audio and visual sensors [4]. It could be used in identifying mobility patterns or popularity of a given trajectory [4]. Such a contribution can be utilized in large-scale urban transit planning [4].

Zhao et al. suggested identifying traffic condition using probe data collected by GPS enabled fleet management devices [9]. By using anonymized signaling data collected from a cellular mobile network Janecek et al. proposed a methodology to infer vehicle travel times and road congestion [10]. D'Andrea and Marcelloni proposed to detect traffic conditions using GPS data provided by smartphone users moving in a road network [11]. The methodology followed in this study is an improved version of the method proposed by D'Andrea and Marcelloni which had the limitation of scalability to a larger network with the available resources [12].

2.3 Use of Machine Learning in Flow Prediction

Use of machine learning principals has become an optimistic approach in traffic flow prediction due to its stochastic and non-linear behavior. In literature, three broader categories of models that are used for traffic flow prediction could be identified as, linear parametric models, non-linear parametric models, and non-linear non-parametric models [1, 13–15]. On the evaluation of linear parametric models, historical average prediction models, time series prediction models, exponential filtering model and Kalman filter model [as cited in 13] are very popular. The researchers who consider the non-linear parametric behavior of traffic flow followed wavelet analysis based models [16], cellular automata model [17], fuzzy regression model [18] and the catastrophe theory-based models [16]. However, limitations in above methods caused poor performance in traffic flow prediction due to linearity and parametric approach. Hence, the researchers have focused on non-linear non-parametric approaches based on machine learning principles. On this aspect, use of k-nearest neighbour's regression model [8, 19], random forest regression model [8], support vector regression model [20], Gaussian process regression model [21] and artificial neural networks based models [1] could be observed.

3 Methodology

In order to predict travel time, the methodology was structured in three stages as data collection, data validation, and analysis. Google Distance Matrix API and the Infra-Red Traffic Logger were used to collect travel time data and traffic flow data. Data collected from both sources were merged together by matching spatiotemporal parameters. The analysis was carried out using the IBM SPSS Modeler machine learning platform [22].

In data collection, traffic flow and travel time data were collected in Sri Lanka at ten locations in the city of Colombo which covered mid blocks (more than 500 m away from intersections) of two-lane, two-way roads. The time interval was set to be 5 min and data collection has conducted continuously in daytime under dry weather condition for three months (Fig. 1).

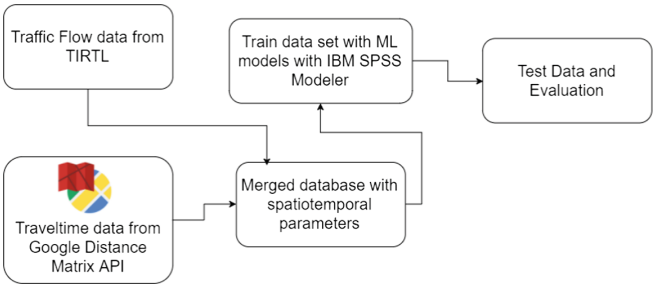

Fig. 1. Flowchart of the methodology.

3.1 Traffic Flow and Travel Time Data Collection and Validation

Traffic flow data was collected through an Infra-Red Traffic Logger (TIRTL) system which is capable of classifying types of vehicles that are defined by axles and vehicle lengths. Video survey has conducted to validate TIRTL data accuracy. A linear regression model was developed to calibrate the instrument [24].

The Google official blog describes, the travel time estimations given in Google Distance Matrix API is based on crowdsourced data [12]. When a smartphone user enables Google services on a smartphone or on a similar device, the application sends anonymous bits of data to Google which carry information on user's location and the moving speed of the user [23]. The travel time prediction algorithms on Google servers combine user's speed with the speeds of other phones on the road, across thousands of phones moving around a city at any given time, in which a significantly reliable estimate on traffic congestion could be gained. This process is continuously carried out to make predictions accurate based on machine learning principles by referring to historical data [12].

Collecting travel time data for flow prediction model was carried out by storing data gathered from API in a cloud server. Google Distance Matrix API was called in one-minute intervals to collect travel time data in between a specific origin and destination

locations where the infrared traffic logger was placed. A self-executing Hypertext Preprocessor (PHP) script was used to utilize the data collection process [27].

The collected data was validated by Floating Car Data (FCD) method which is a manual data collection using GPS device equipped vehicles moving in traffic flow. Travel time data collected was then compared with Google Distance Matrix API travel time data [27]. The observed variation is within 5%–15% range which indicates that the travel time estimates given by Google Distance Matrix API have no significant difference from floating car data (FCD). Figure 2 shows a comparison of travel time data obtained from the above mentioned two methods for 108 events in Colombo urban road network [27].

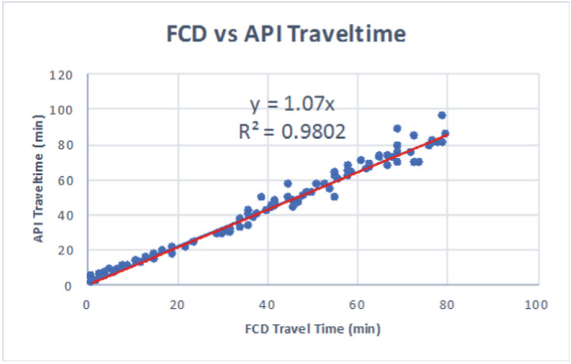

Fig. 2. Comparison of API travel time with FCD travel time Ref: [27].

3.2 Evaluation Using Machine Learning

In this research, traffic flow is considered as a non-linear non-parametric function of temporal inputs and spatial inputs (see Table 1). A clustering based regression method is proposed by following the objective of the study. Zhang et al. and Zhong et al. have concluded that a relatively accurate model can be obtained with an average sized dataset when K- Nearest Neighbours (KNN) is used [25].

KNN is a non-parametric algorithm which model parameters do not have to be calculated. K is the number of nearest neighbours considered in classification. When K is defined, the prediction is done by identifying the K nearest data points and counting

Table 1. KNN input attributes.

Temporal attributes (T)	Spatial attributes (S)
t = Time (hh:mm)	l = Link Length (km)
s = Link Speed (km/h)	w = Lane Width (m)
s_o = Opposite Link Speed (km/h)	w_o = Opposite Lane Width (m)
	h = Shoulder Width (m)
	h_o = Opposite Shoulder Width (m)

the frequency of each class among the K nearest neighbours. The K value should be defined in order to minimize the absolute error of the model [26].

In this study, the processed dataset was analyzed using the KNN model supported by IBM SPSS Modeler [22]. Eight spatiotemporal attributes were considered in flow prediction as given in Table 1. The Eq. 2 illustrates the flow prediction concept used in KNN algorithm. When the model is trained, it was verified using the test dataset for prediction accuracy. The Eq. 3 illustrates the use of distance weights in obtaining higher accuracy [26].

$$Flow_{KNN} = f\left(T(t,s,s_o)_{1,2,\ldots..n}, S(l,w,w_o,h,h_o)_{1,2,\ldots..n}\right) \tag{2}$$

$$Flow_{predict} = \frac{1}{K}\sum_{i=1}^{K} w_i Flow_{KNN} \tag{3}$$

The main evaluators of the machine learning models are the Root Mean Squared Error (RMSE) and the Maximum Absolute Error (MAE). In order to compare the results obtained the same dataset was trained under support vector regression (SVR) and artificial neural networks (ANN).

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x_l})^2} \tag{4}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\frac{|x_i - \overline{x_i}|}{x_l} \tag{5}$$

Where, x_i is true value, $\overline{x_l}$ is predicted value and N is number of patterns.

4 Analysis

4.1 Results and Discussion

A 70% of the dataset was used to train the model while 30% was utilized for testing the trained model. Figure 3(right) shows the distribution of training dataset and the test dataset.

K- Nearest Neighbours (KNN) regression was carried out with K = 3 as it shows the minimum sum of squares error when compared with K greater than 3 (Fig. 3-Left). Further KNN regression with K = 3 gives the minimum prediction errors since RMSE and MAE is lower when K = 3 than K = 4 or K = 5 (see Table 2).

Compared to K- Nearest Neighbour regression the artificial neural network and support vector regression has given higher prediction errors (see Table 2).

The Fig. 4 shows the variation between predicted flow data and actual flow data obtained from neural network regression (Fig. 4-left) and support vector regression (Fig. 4-right). The distribution of predicted data does not match with the distribution of original dataset in both neural network and support vector regression.

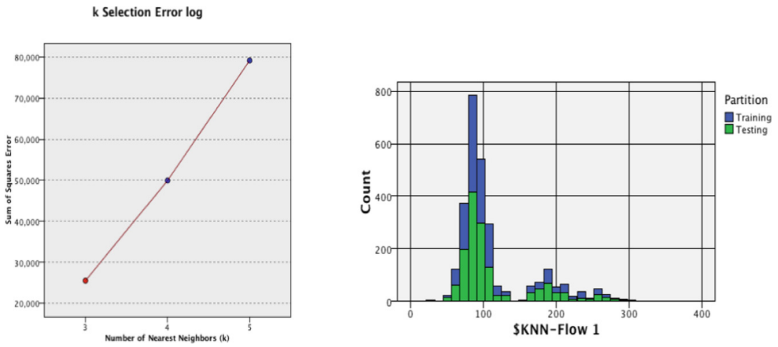

Fig. 3. Left: K- selection error graph; Right: Training and test sample distribution.

Table 2. Evaluation of regression models.

Parameter	RMSE	MAE	Linear correlation
KNN k = 3	9.479	2.318	0.983
KNN k = 4	12.762	3.67	0.962
KNN k = 5	15.452	4.237	0.924
SVR	44.691	25.854	0.528
ANN	29.13	18.31	0.710

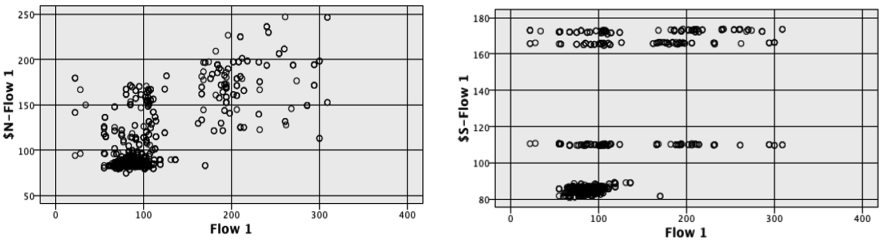

Fig. 4. Predicted Vs Observed traffic flow for ANN (left) and SVR (right).

Figure 5 shows the comparison of predicted values with actual values of the directional flow when plotted against time. The predictions are much accurate with the K- Nearest Neighbour analysis when compared the linear correlation (see Table 2).

When considered the input parameters which contribute to flow prediction the directional link speed and opposite direction has the major influence. Figure 6 shows the scatter of directional speeds with predicted Flow values which depict the non-linear behavior.

The verification of the model by cross-validation is conducted using n-fold validation by splitting the dataset into 12 folds. The agreement between predicted KNN-Flow vs actual Flow for each fold was analyzed by taking the RMSE and MAE of

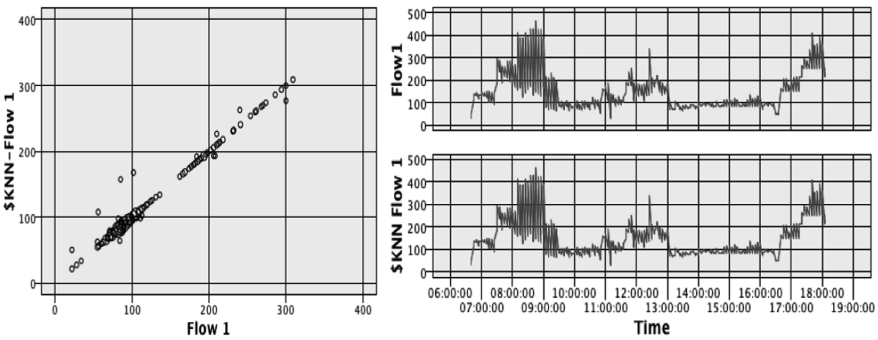

Fig. 5. Predicted and observed flow values under KNN regression.

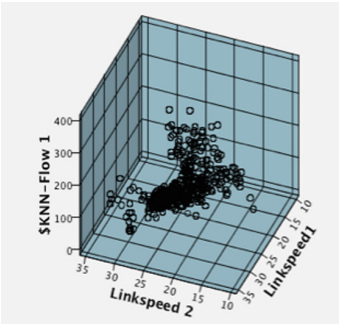

Fig. 6. Distribution of directional and opposite link speeds with predicted flow.

Table 3. Evaluation for model over fit.

Partition	RMSE	MAE	Linear correlation
Training	11.504	5.91	0.979
Testing	11.92	6.445	0.976
% Difference	3.49%	8.3%	–

training dataset and testing datasets for each fold. Table 3 gives the average RMSE and MAE obtained by averaging each fold RMSE and MAE.

As the percentage difference of RMSE and MAE are at low values it can be concluded that the model is not overfitting. The Fig. 7 shows the percentage gain of each training dataset which indicate that the model could be validated for better results.

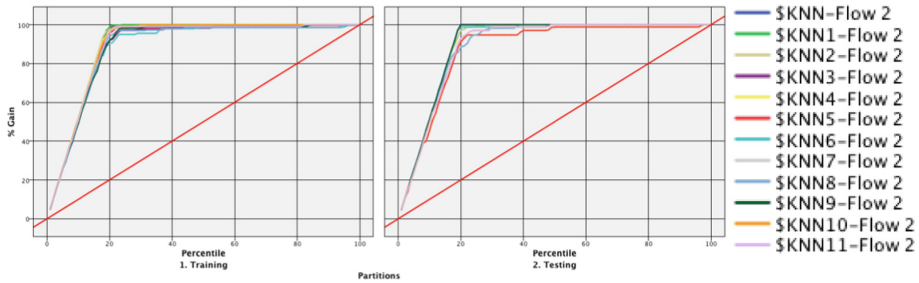

Fig. 7. The percentage gain of each training dataset.

5 Conclusion

The results show that traffic flow prediction for urban two-lane roads using machine learning principles become a success. Deviating from traditional methods of linear parametric approaches for flow prediction the study evaluates the use of the non-linear non-parametric approach. The study develops a traffic flow estimation model based on K- Nearest Neighbour regression which uses spatiotemporal inputs. Travel time and speed data obtained from Google Distance Matrix API and road geometry data are used as inputs to the model.

K- Nearest Neighbour regression was able to give a higher prediction accuracy with a linear correlation of 0.97. The prediction error was minimum at K = 3 neighbours. The model was cross-validated using the N-Fold cross-validation method and showed the model does not over fit.

In future work, it is suggested to incorporate continuous data feeding methods. Further incorporating data sources such as terrain details, pavement conditions, and land use as categorical variables could enable the estimation model being used at rural areas.

The success of the study enables to use Google Distance Matrix API travel time data in estimating the traffic flow which is an economic alternative for developing countries.

References

1. Dai, X., Fu, R., Lin, Y., et al.: DeepTrend: A Deep Hierarchical Neural Network for Traffic Flow Prediction (2017)
2. Japan International Cooperation Agency; Oriental Consultants Co., LTD. Urban Transport System Development Project For Colombo Metropolitan Region
3. Amini, S., Gerostathopoulos, I., Prehofer, C.: Big Data Analytics Architecture for Real-Time Traffic Control
4. Chatzimilioudis, G., Konstantinidis, A., Laoudias, C., Zeinalipour-yazti, D.: Crowdsourcing with smartphones. IEEE Internet Comput. **16**(5), 1–7 (2012). https://doi.org/10.1109/MIC.2012.70

5. Russell, R.: How does Google maps calculate your ETA. In: Forbes (2013). https://www.forbes.com/sites/quora/2013/07/31/how-does-google-maps-calculate-your-eta/#241f6c01466e
6. Helbing, D.: From microscopic to macroscopic traffic models. In: Parisi, J., Müller, S.C., Zimmermann, W. (eds.) A Perspect. Look Non-linear Media, vol. 503, pp. 122–139. Springer, Heidelberg (2012). https://doi.org/10.1007/BFb0104959
7. Chandra, S.: Capacity estimation procedure for two-lane roads under mixed traffic conditions. J. Indian Roads Congr. **i**, 139–167 (2004)
8. Antoniou, C., Koutsopoulos, H.: Estimation of traffic dynamics models with machine-learning methods. Transp. Res. Rec. J. Transp. Res. Board **1965**, 103–111 (2006). https://doi.org/10.3141/1965-11
9. Zhao, W., McCormack, E., Dailey, D.J., Scharnhorst, E.: Using truck probe GPS data to identify and rank roadway bottlenecks. J Transp. Eng. **139**, 1–8 (2013). https://doi.org/10.1061/(ASCE)TE.1943-5436.0000444
10. Janecek, A., Hummel, KA., Valerio, D., et al.: Cellular data meet vehicular traffic theory: location area updates and cell transitions for travel time estimation. In: ACM Conference on Ubiquitous Computing, pp. 361–370 (2012)
11. D'Andrea, E., Marcelloni, F.: Detection of traffic congestion and incidents from GPS trace analysis. Expert Syst. Appl. **73**, 43–56 (2017). https://doi.org/10.1016/j.eswa.2016.12.018
12. Google. The bright side of sitting in traffic: Crowdsourcing road congestion data. Googleblog (2009)
13. Cheng, A., Jiang, X., Li, Y., et al.: Multiple sources and multiple measures based traffic flow prediction using the chaos theory and support vector regression method. Phys. A Stat. Mech. Appl. **466**, 422–434 (2017). https://doi.org/10.1016/j.physa.2016.09.041
14. Laboshin, L.U., Lukashin, A.A., Zaborovsky, V.S.: The Big Data approach to collecting and analyzing traffic data in large scale networks. Procedia Comput. Sci. **103**, 536–542 (2017). https://doi.org/10.1016/j.procs.2017.01.048
15. Xu, C., Li, Z., Wang, W.: Short-term traffic flow prediction using a methodology based on autoregressive integrated moving average and genetic programming. Transport **31**, 343–358 (2016). https://doi.org/10.3846/16484142.2016.1212734
16. Elsner, J.B., Tsonis, A.A.: Non-linear Prediction, Chaos, and Noise. Bull. Am. Meteorol. Soc. **73**, 49–60 (1992). https://doi.org/10.1175/1520-0477(1992)0732.0.CO;2
17. Bao, J., Chen, W., Xiang, Z.: Prediction of traffic flow based on cellular automaton. In: 2015 International Conference on Industrial Informatics - Computing Technology, Intelligent Technology, Industrial Information Integration, pp. 88–92 (2015). https://doi.org/10.1109/iciicii.2015.107
18. Shang, Q., Lin, C., Yang, Z., et al.: A hybrid short-term traffic flow prediction model based on singular spectrum analysis and kernel extreme learning machine. PLoS ONE **11**, 1–25 (2016). https://doi.org/10.1371/journal.pone.0161259
19. Zhang, L., Liu, Q., Yang, W., et al.: An improved K-nearest neighbour model for short-term traffic flow prediction. Procedia – Soc. Behav. Sci. **96**, 653–662 (2013). https://doi.org/10.1016/j.sbspro.2013.08.076
20. Castro-Neto, M., Jeong, Y.-S., Jeong, M.-K., Han, L.D.: Online-SVR for short-term traffic flow prediction under typical and atypical traffic conditions. Expert Syst. Appl. **36**, 6164–6173 (2009). https://doi.org/10.1016/j.eswa.2008.07.069
21. Zhao, J., Sun, S.: High-order Gaussian process dynamical models for traffic flow prediction. IEEE Trans. Intell. Transp. Syst. **17**, 2014–2019 (2016). https://doi.org/10.1109/TITS.2016.2515105
22. IBM Corp. IBM SPSS Modeler for Windows. (2016)

23. Gunter, U., Onder, I.: Forecasting city arrivals with Google Analytics. Ann. Tour Res. **61**, 199–212 (2016). https://doi.org/10.1016/j.annals.2016.10.007
24. Rajapaksha, R.P.G.K.S., Bandara, J.M.S.J.: Effect of traffic composition on capacity of two-way two-lane, roads under mix traffic condition. In: International Conference on Advances in Highway Engineering & Transportation Systems, vol. 20 (2017)
25. Zhong, J., Ling, S.: Key factors of k-nearest neighbours nonparametric regression in short-time traffic flow forecasting. In: Qi, E., Shen, J., Dou, R. (eds.) Proceedings of the 21st International Conference on Industrial Engineering and Engineering Management 2014. PICIEEM, pp. 9–12. Atlantis Press, Paris (2015). https://doi.org/10.2991/978-94-6239-102-4_2
26. Wendler, T., Gröttrup, S.: Data Mining with SPSS Modeler. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-28709-6
27. Kumarage, S.P., De Silva, D., Bandara, J.M.S.J.: Travel time estimation based on dynamic traffic data and machine learning principles. In: IESE Annual Sessions 2017, pp. 1135–1142 (2017)