Analyzing the Accuracy of Historical Average for Urban Traffic Forecasting using Google Maps

Hajar Rezzouqi, Ihsane Gryech, Nada Sbihi, Mounir Ghogho, Houda Benbrahim

Intelligent Systems Conference 2018
6-7 September 2018 | London, UK

Analyzing the Accuracy of Historical Average for Urban Traffic Forecasting using Google Maps

Hajar Rezzouqi[1,2], Ihsane Gryech[1,2], Nada Sbihi[1], Mounir Ghogho[1,3], Houda Benbrahim [2]

[1]Université Internationale de Rabat, FIL, TICLab, Morocco
[2]IRDA, Rabat IT Center, ENSIAS, Université Mohammed V, Rabat, Morocco
[3]The School of EEE, the University of Leeds, UK

{hajar.rezzouqi, ihsane.gryech, nada.sbihi, mounir.ghogho}@uir.ac.ma
houda.benbrahim@um5.ac.ma

***Abstract*—Urban traffic management uses increasingly sophisticated methods to overcome the many challenges involved in the development of traffic forecasting solutions. The main challenge is the acquisition of real-time large-scale urban traffic data at a sufficient spatio-temporal resolution. This is a challenge mainly because of the high financial cost that the installation of a large number of sensors would incur. This paper addresses this challenge by leveraging 'real-time' Google Traffic maps which show the state of the traffic on different road segments using four different colors. Since Google Traffic maps are provided in the form of rendered images, we apply image processing on Google Traffic maps to extract traffic data that are suitable for processing and analysis. The traffic data obtained from Google Traffic are validated with a traffic data set collected by sensors installed in Paris. Then, using data gathered through our Google Traffic-based method for several roads in Rabat (Morocco), we evaluate the accuracy of traffic prediction based on historical average. The overall accuracy reaches 74.9% on week days and 83.3% on weekend days. Further, a more detailed study by type of road and by time period was conducted, showing an overall accuracy of 95.8% in fluid traffic situations.**

***Keywords*—Google traffic; image processing; data collection; prediction; historical average**

I. Introduction

The expanding traffic flow in cities could lead to a transport crisis. There are several dimensions that are affected by urban traffic, but the most negatively impacted ones are the environment and the economy. Therefore, optimising and predicting urban traffic are necessary to monitor and control congestion and pollution, and are thus at the core of research and development in smart cities. This field of research requires an understanding and analysis of traffic behaviour, hence the need for massive traffic data for the many different roads of the city.

There are various methods used for the acquisition and analysis of traffic data. Regarding the data acquisition, the most commonly mentioned sensors in the research literature are: inductive loop vehicle detectors, cameras, vehicles equipped with GPS, smartphones, and surveys that can be used to estimate the origin-destination matrices.

In [1], the collection of the individual itineraries was carried out through the use of vehicles equipped with data loggers. In [2] and[3], taxis equipped with GPS, considered as mobile sensors, provide large-scale real-time traffic traces which allow for a global view of the dynamics of the urban road network. The authors in [4] argued that the smartphone can be considered as a reliable source for data acquisition. A correlation coefficient was computed between the generated variables from the on-board instrumentation and those provided from a smartphone equipped with a GPS receiver, a triaxial accelerometer, a triaxial gyroscopic and a compass, yielding results between 99% and 100%. Data collection can also be accomplished by more advanced sensors such as the Electronic Traffic Bayonet Device (ETBD). The latter enables to take pictures of each passing vehicle, automatically identifies the license plate, the vehicle brand, the speed, etc. and then stores the information in a database [5]. Some other new technologies have also been deployed to record road traffic data. For example, in [6], Bluetooth was proven to be easy to install and maintain and does not raise privacy concerns; the MAC address and the exact time of detection of the individual devices detected by the Bluetooth sensor were recorded. It was then possible to estimate the travel time between two different detections and hence speed of the vehicle.

Even though the technologies allowing traffic data collection are in constant evolution, ensuring such collection in urban areas faces obstacles which hinder urban traffic analysis. Indeed, efficient collection requires a massive and costly installation of sensors. This is particularly challenging for developing countries. To address the above-mentioned issue, we propose a novel method of traffic data collection based on Google Traffic, which is a feature displayed on Google Maps to indicate the level of congestion on roads. We evaluate the validity of Google Traffic data using an open traffic dataset from the city of Paris. After this validation step, we address the problem of predicting the level of congestion on some roads in the city of Rabat using Google Traffic.

II. Methodology

The scarcity of traffic sensors on Moroccan roads makes traffic prediction a difficult task. Smartphones represent a reliable source of information about some traffic parameters. Unfortunately, getting hold of smartphone data for researchers is difficult as mobile operators are reluctant to share such

978-1-5386-4185-9/18/$31.00 ©2018 IEEE

data because of privacy concerns. To circumvent this problem, we propose to extract information from Google Traffic maps, which show the level of congestion on the main roads of most cities. Indeed, Google estimates the level of congestion on a road by analysing the GPS data obtained from the Android smartphones carried by the vehicles passing through the road. The level of congestion is determined by the average speed on the road over a given time period [7]. As shown in Fig. 1, for each congestion level, there is a specific color: the green color represents fluid traffic, orange color indicates less traffic congestion, red denotes congestion while brown refers to high traffic congestion. When smartphone data are not available, none of these colors is shown on the corresponding roads. Therefore, the reliability of Google Traffic improves with the penetration of android smartphones.

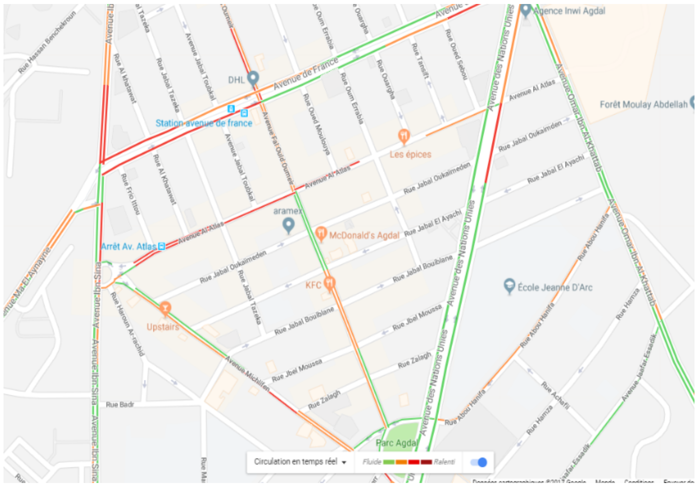


Fig. 1. Congestion levels of an area of Rabat city displayed on a Google Traffic map.

The methodology that we have adopted in our work to extract and analyse the data obtained from Google Traffic maps is based on the implementation of the following steps (as described in Fig. 2):

- Automatic maps collection: The area and roads of interest are first selected; then, an automatic screen capture of the corresponding Google map (with traffic feature on) is programmed to obtain a traffic map every 5 minutes; this is accompanied by an automatic refresh of the page.
- Image processing: Applying basic image processing tools on the built image database, relevant traffic information is extracted every 5 minutes.
- Data processing: This consists of building the final database where each data element consists of the road identity, the date and time of capture, and the level of congestion.
- Traffic prediction: Since the state of the traffic is represented by a categorical variable (4 levels of congestion), the prediction is cast as a classification task. In this paper, we limit our traffic prediction study to historical average.
- Performance analysis.

A. Image Processing

The image processing consists of extracting the RGB code of the positions p representing the coordinates of the pixels chosen to represent the different roads of interest. These positions were manually determined before the automatic data collection. Each position 'pixel' has its own RGB code. Since for each color we have different shades, we were faced with a multitude of RGB codes. For this reason and to facilitate our data analysis, we decided to use as centroids of the RGB codes the main colors needed, namely, green, orange, red, brown and white (which indicates that no data is available). Once the centroids have been fixed, the Euclidean distances between the RGB codes and these centroids are calculated, and the minimum distance determines the color category the RGB code belongs to. Thus, this distance is given by

$$d(p,c_j) = \sqrt[2]{(R_p - R_{c_j})^2 + (G_p - G_{c_j})^2 + (B_p - B_{c_j})^2}$$

where p is a specific position on the map whose extracted RGB code is represented by the tuple (R_p,G_p,B_p), and the tuple (R_{c_j},G_{c_j},B_{c_j}) represent the RGB code of the jth centroid (j=1,...,5).

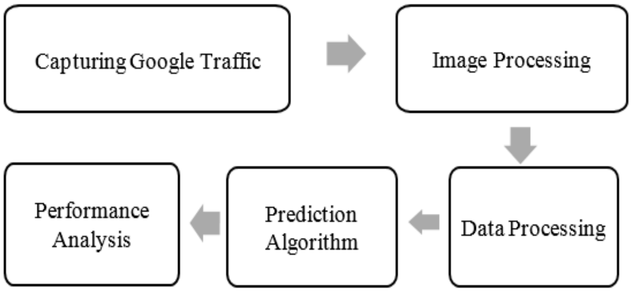


Fig. 2. Traffic data collection and analysis process.

B. Validation of the Proposed Data Collection Method: A Comparative Study

To test the proposed approach of collecting traffic data from Google Maps, and to check the validity of our extracted data, we used an existing dataset from the city of Paris [8]. This

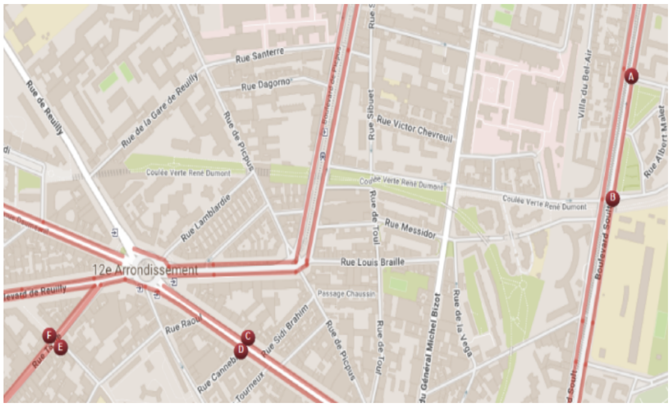


Fig. 3. Congestion levels of an area of Rabat city displayed on a Google Traffic map.

publicly available dataset is provided by the city hall of Paris, and collected using permanent traffic sensors installed on the urban network of the city. Only hourly traffic data is available. The dataset includes:

- The flow: the number of vehicles that passed by the measuring point during a time period (one hour).
- The occupancy rate: this corresponds to the time of presence of vehicles on a loop, as a percentage of a fixed time period.

Fig. 3 illustrates the studied area. We next compare our newly collected dataset, and the dataset provided by Open Data Paris over the same period of time. To do so, we considered two separate measurement points (A and B) on the large “Boulvard Soult” road, two measurement points (E and F) on the small “Rue Traine” road, one on each lane, and two other measurement points (C and D) on “Rue Cannebire”. The selected day chosen for the comparative study was 17 October 2017. With the data collected from Google about the small roads (E, F), we obtained a green shade during almost the whole period of extraction. Thus, there is an absence of variation in the level of congestion unlike the data collected from the installed sensors. This makes it hard to match the level of congestion with traffic parameters such as the flow, and the occupancy ratio. However, for large roads (A,B), we were able to detect this variation.

There is also an issue with the time resolution of the traffic measurements in the two datasets: the screen shots from Google Traffic are taken every 5 minutes, and the sensor measurements are provided hourly. Within each hour, each 5 minutes period is associated with one level of congestion. Hence, in order to be able to compare the two sets of measurements, we aggregate Google traffic data using one of the following methods.

The first method consists of calculating the mode during each hour, i.e. the predominant level of congestion. The limitation of this method is that the dominant level is very often green, which takes us back to the problem we faced with the two points of measurement (E,F). This is illustrated in Fig. 4 where the blue curve corresponds to the level of congestion retrieved from google, while the green curve describes the data retrieved from the Paris dataset.

The second method assumes that the level of congestion representing any given hour is the corresponding highest level of congestion. By selecting the maximum level of congestion, we observe that the traffic flow, the occupancy rate and the level of congestion curves vary in the same way (Fig. 5 and 6). This implies that the data extracted from Google Maps are in agreement with those obtained from the sensors and thus reflect the real state of traffic.

C. Disaggregation of traffic data

The traffic parameters available in the Paris dataset are measured hourly. One may be interested in disaggregating these measurements in order to understand the dynamics of the traffic at a higher temporal resolution. Further, we also notice that there are sometimes many missing values which require to be imputed. To address the above mentioned issues, we propose to combine the hourly sensor data with Google Traffic maps. More specifically, we derive a formula to extract traffic values (Number of vehicles, occupancy rate, average speed) at a five minute-temporal resolution from the Google Maps congestion levels and the hourly data provided in the Paris dataset. Consider the average speed as an example. Let X be the average speed recorded per hour, and let X_i be the average speed over the ith five minute period. We would like to estimate the X_i's using X and the congestion levels that are collected from Google Traffic every five minutes. We propose to estimate the X_i's using the following method:

$$X_i = w_i.X, \quad i = 1, ..., 12 \tag{1}$$

subject to

$$X = \frac{\Sigma_{i=1}^{12} X_i}{12}, \qquad \Sigma_{i=1}^{12} w_i = 12 \tag{2}$$

where the weights w_i are to be determined using the levels of congestion extracted from Google Traffic maps. Indeed, the more congested a road is, the lower the average speed. This ‘proportionality’ could be leveraged to compute the weights. There is however no trustworthy reference regarding the average speed thresholds on which Google Traffic rely to define the four levels of congestion. This implies that there will always be an uncertainty about the weight calculation.

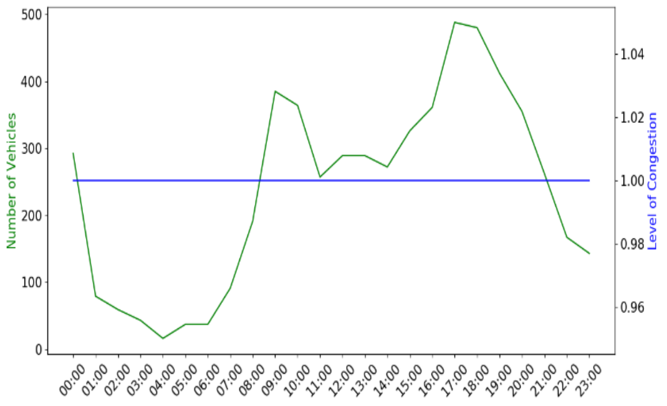


Fig. 4. Comparison between Paris Data set and Google Traffic Data set based on the mode (Measurement point: B).

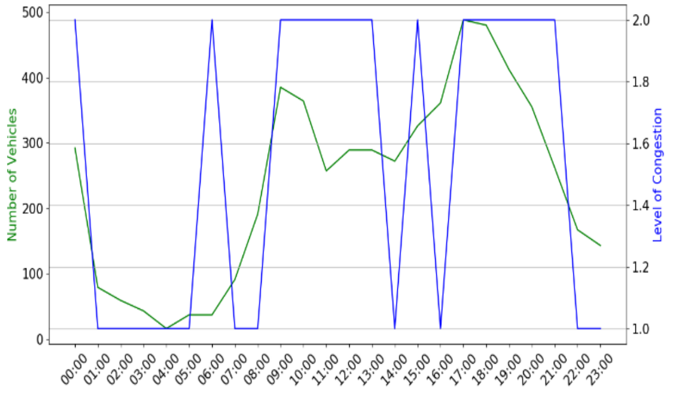


Fig. 5. Comparison between Paris Data set and Google Traffic Data set based on the highest congestion level (Measurement point: B).

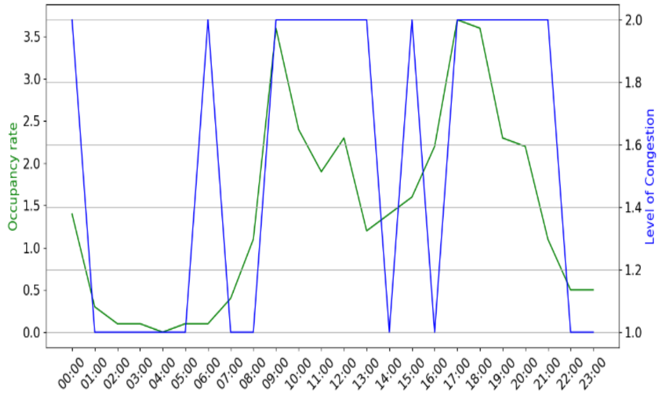
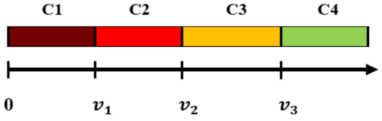


Fig. 6. Comparison between the occupancy rate extracted from Paris Data set and Google Traffic Data set based on the highest congestion level (Measurement point: B).

Nevertheless, we here make the assumption that the average speed thresholds, v_1, v_2, v_3 and v_4, defining the different levels of congestion are equidistant, i.e.:

$$v_3 - v_2 \simeq v_2 - v_1 \simeq v_1 - 0 \implies v_2 \simeq 2.v_1, \quad v_3 \simeq 3.v_1$$



Hence, the weights can be estimated as follows:

$$w_i = 12\frac{z_i}{\Sigma_{i=1}^{12} z_i} \qquad i = 1, .., 12 \tag{3}$$

where z_i, which takes on values from the set $\{1, 2, 3, 4\}$, is determined from the Google Traffic-derived congestion level associated with the ith five minute period; $z_i = 4$ for brown color, i.e. very congested traffic, $z_i = 3$ for red, $z_i = 2$ for orange and $z_i = 1$ for green.

III. Prediction Model

In this paper, traffic prediction is merely based on the historical averages of the data extracted from the images obtained with Google Traffic. The database obtained through image processing consists of the date/time of the screen capture, the different positions corresponding to the selected roads, and the colors that reflect the level of congestion.

A white color on the map means that either there are no vehicles on the corresponding road, or there are no smartphones detected. We also noticed that this color is the most common color at night. This may introduce a bias to our prediction model. So, in our study, we assumed that most road users are equipped with at least one smartphone, and thus not detecting a smartphone means very few vehicles passing through the road and hence a fluid traffic (green color). As a result, we merged the white color with the green one.

We have selected a specific area of Rabat city to test our approach. We have collected Google Traffic data from 22 September 2017 at 00:05 to 23 October 2017 at 23:55, i.e. over a period of 32 days, with a frequency of one map per 5 minutes. 49 positions were chosen for this study, giving a database of 451,535 observations. Since our prediction model is based on historical average, it assumes that the traffic has a strong periodic component, with may change with the type of road and time period. For this reason, we have distinguished week days from weekend days, and grouped the data by road and subsequently by hour. Once the data are aggregated to ensure the same spatial and temporal conditions, the mode is determined. Our database was divided into two parts: training and testing sets. For traffic prediction on a test week day, all the data corresponding to the week days are aggregated from the training data, and the level of congestion for the test day at time t is estimated according to the following formula:

$$l_{t,i+1} = \text{Mode}(l_{t,i}, l_{t,i-1}, l_{t,i-2}, ..., l_{t,i-n}) \tag{4}$$

where $l_{t,i+1}$ denotes the level of congestion to be predicted on weekday $i+1$ at time t, and $n+1$ corresponds to the number of days used in the training data set.

For predicting the traffic on weekends, the same approach as above is applied using the training data corresponding to weekend days.

IV. Performance Analysis

The performance of the prediction of the level of congestion on the selected roads is evaluated using the following metrics: Accuracy, Precision, Recall and F1 score. The results corresponding to the prediction of traffic on week days and weekend days are presented in Tables I and II, respectively.

TABLE I. The Performance of the Historical Average-Based Prediction for Week Days

	The Historical Average
Accuracy	74.9%
Precision	71.9%
Recall	74.9%
F1	72.7%

TABLE II. The Performance of the Historical Average on Weekends

Accuracy	83.3%
Precision	78.5%
Recall	83.3%
F1	80%

There are several reasons for traffic congestion, the most important of which road construction work, accidents, interruptions, delays caused by traffic lights, stopping points, and intersections. Indeed, with the high number of cars on a road, any traffic disruption can cause a traffic jam. In some cases, bottling is not directly or rationally justified. This is mainly due to the individual behaviour of drivers, especially those who drive very aggressively or those who drive very carefully and slow down traffic. Thus, a model that relies on the historical

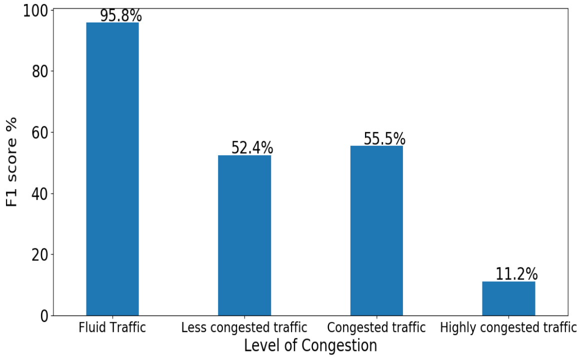


Fig. 7. Global performance for each level of congestion.

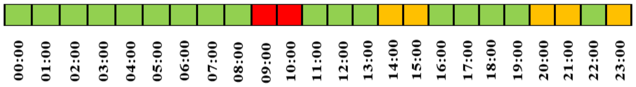


Fig. 8. Level of congestion per time period during weekdays.

average is not unable to predict these unexpected occurrences, since it only takes into account the usual conditions on any given route. This may justify the overall accuracy obtained in the paper: 74.9% on weekdays and 83.3% on weekends (Tables I and II); this is also illustrated in Fig. 7.

Further, we have examined the prediction performance related to each congestion level, and found that the more fluid the traffic is, the better the F1 score of the historical average, which reaches 95.8%. On the other hand, the more congested traffic becomes, and this is quite likely due to the above-mentioned causes, the lower the F1 score, which may be as low as 11.2%.

We also studied the performance of the historical average per time period during week days (see Fig. 10). The results are coherent with what has been found regarding the performance by congestion level: during peak hours, when traffic is congested (Fig. 8 and 9), the accuracy becomes less satisfactory (e.g. the overall accuracy at 10:00am is 52%), whereas in the hours where the traffic is fluid (for example at 4am and 5am), a high precision has been reached (nearly 100%). A similar approach has been applied to weekends, (Fig. 11) but here the peak hours are not the same as on week days (Fig. 9). Indeed, people tend to go out more in the evenings between 07:00pm and 00:00am. This explains the lower performance during these hours and a better performance in the very early morning hours.

Finally, we have found that for roads with a more dynamic traffic (presence of shopping centres, schools, companies, etc.), the accuracy of the historical average-based prediction at peak hours decreases. Indeed, it is in these hours that we notice rapid fluctuations in the level of congestion, which makes it difficult to predict the level of congestion at a specific hour, based only on the historical average.

Traffic prediction using the historical average is characterised by ease of implementation and speed of execution. The more traffic data is available, the better the performance of this method gets. This method however may not be sufficiently accurate in the presence of unexpected events [9]. The other limitation is that the level of congestion at any given time is estimated based on the history of traffic over the last few days, and thus does not take into account the traffic conditions of the immediate past (last minutes or hours). Other factors influencing the traffic such as the weather have not been taken into account in our model. We are currently investigating this issue as well as exploring the spatio-temporal relationships between roads in order to increase the accuracy of prediction.

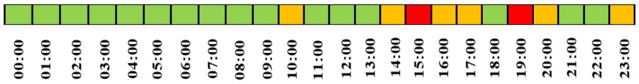


Fig. 9. Level of congestion per time period during weekends.

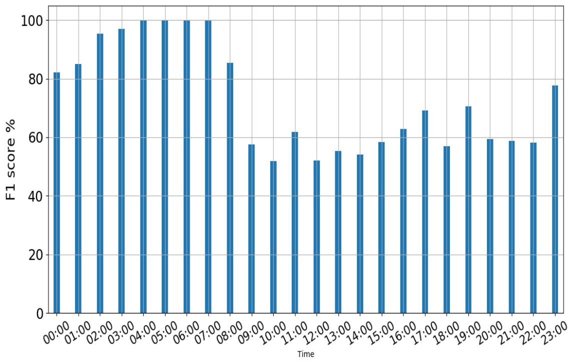


Fig. 10. Global performance by time period during weekdays.

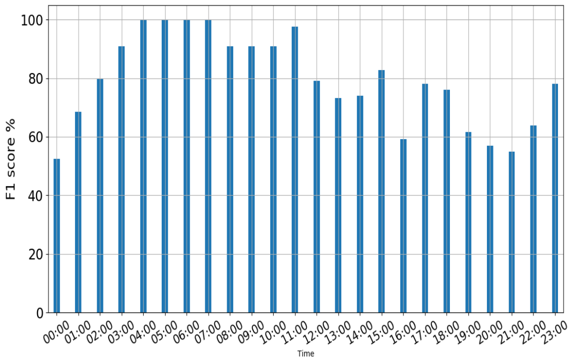


Fig. 11. Global performance by time period during weekends.

V. Conclusion and Future Work

We have proposed a novel method of urban traffic data collection which consists of exploiting Google Traffic Maps using image processing. A validation method, based on a comparative study with an existing data set from the city of Paris, was proposed to verify the reliability of the extracted Google Traffic data. This study demonstrated that there is a correlation between the level of congestion obtained from Google Traffic maps and the number of vehicles on roads, as well as the occupancy rate retrieved through the sensors installed in Paris. We have also proposed a way to disaggregate traffic data obtained with sensors using Google Traffic maps. Using the proposed traffic data collection, we built a database about the traffic in an urban area of the the city of Rabat (Morocco). Using this dataset, we have evaluated the performance of the historical average-based traffic prediction of the level of traffic congestion. Week and weekend days were treated separately. The perceived limitations of this method, due to unexpected events and changing weather conditions are motivations for future work.

Acknowledgment

This work is a part of the MOROCCAN VIDEO INTELLIGENT TRANSPORT SYSTEM project, and is funded by "Centre National pour la Recherche Scientifique et Technique" (CNRST).

References

[1] Park, J., Murphey, Y. L., Kristinsson, J., McGee, R., Kuang, M., & Phillips, T. (2013). Intelligent speed profile prediction on urban traffic networks with machine learning. Proceedings of the International Joint Conference on Neural Networks.

[2] Yu, B., Song, X., Guan, F., Yang, Z., & Yao, B. (2016). k-nearest neighbor model for multiple-time-step prediction of short-term traffic condition. Journal of Transportation Engineering, 142(6), 1-10.

[3] Deng, B., Denman, S., Zachariadis, V., & Jin, Y. (2015). Estimating traffic delays and network speeds from low-frequency GPS taxis traces for urban transport modelling. European Journal of Transport and Infrastructure Research, 15(4), 639-661.

[4] Diaz Alvarez, A., Serradilla Garcia, F., Naranjo, J. E., Anaya, J. J., & Jimenez, F. (2014). Modeling the driving behavior of electric vehicles using smartphones and neural networks. IEEE Intelligent Transportation Systems Magazine, 6(3), 44-53.

[5] Yuan, W., Deng, P., Taleb, T., Wan, J., & Bi, C. (2016). An Unlicensed Taxi Identification Model Based on Big Data Analysis. IEEE Transactions on Intelligent Transportation Systems, 17(6), 1703-1713.

[6] Park, H., & Haghani, A. (2015). Optimal number and location of Bluetooth sensors considering stochastic travel time prediction. Transportation Research Part C: Emerging Technologies, 55, 203-216.

[7] Stenovec, T. (n.d.). How Google Maps knows about traffic - Business Insider. Retrieved December 6, 2017, from http://www.businessinsider.com/how-google-maps-knows-about-traffic-2015-11?IR=T

[8] Open Data Paris. (2011). Open Data Paris. Retrieved December 8, 2017, from https://opendata.paris.fr/page/home/

[9] Bolshinsky, E., & Freidman, R. (2012). Traffic Flow Forecast Survey, 1-15.