Understanding Daily Mobility Patterns in Urban Road Networks using Traffic Flow Analytics

Ibai Laña*, Javier Del Ser*,† and Ignacio (Iñaki) Olabarrieta*
*TECNALIA RESEARCH & INNOVATION, 48170 Derio, Bizkaia, Spain
Email: {ibai.lana, javier.delser, ignacio.olabarrieta}@tecnalia.com
†University of the Basque Country UPV/EHU, Bilbao, Bizkaia, Spain
Email: javier.delser@ehu.eus

***Abstract*—The MoveUs project funded by the European Commission aims to foster sustainable eco-friendly mobility habits in cities. In this context predicting the traffic flow is useful for managers to optimize the configuration of the road network towards reducing the congestions and ultimately, the pollution. With the explosion of the so-called Big Data concept and its application to traffic data, a wide range of traffic flow prediction methods has been reported in the related literature. However, most of the efforts in this field have been hitherto focused on short-term prediction models. This paper analyzes how to properly characterize traffic flow in urban road scenarios with an emphasis on the long term. To this end a clustering stage is utilized to discover typicalities or patterns within the traffic flow data registered by each road sensor, which permits building prediction models for each of such discovered patterns. These individual prediction models are intended to become part of the MoveUs platform, which will provide the technical means 1) for traffic managers to analyze in depth the status of the road network, and 2) for road users to better plan their trips.**

***Index Terms*—Long-term traffic flow prediction, mobility patterns, machine learning.**

I. Introduction and related work

Traffic flow prediction models help in the development of Intelligent Transportation Systems (ITSs), where they play a role in aiding road users to plan their trips with rich predictive information about the status of the road and different options for public transport. They are also useful for traffic managers, who can improve the road features and configuration, to alleviate congestion or reduce traffic incidences in advance, by predicting the future states of the road at hand [1].

The growing interest in this area, combined with the also increasing portfolio of techniques and technologies for data analysis, have lead to a variety of methods for analyzing traffic data and designing prediction models. Among them Autoregressive Integrated Moving Average (ARIMA) models have been used since the late 1970s to predict short-term traffic flows [2]. Several works have thereafter introduced variations to off-the-shelf ARIMA models to perform time-based predictions, from seasonal ARIMA [3], [4] to hybrid models incorporating Kohonen Maps [5]. Other approaches have instead gravitated on nonparametric methods and schemes from machine learning. As such, Artificial Neural Networks (ANN) have been used extensively [6], [7], [8], [9], [10] with different settings and varying results. Support Vector Regression (SVR) is another method utilized for the same purpose [11], which has been shown to outperform their neural network counterparts [12]. The aforementioned techniques are combined by some researchers to obtain hybrid models, e.g. neural networks have been utilized to improve the tuning procedure of the ARIMA model parameters [13], [14]. In general, statistical techniques with varying tunning parameters outperform nonparametric methods [16], but the latter have the advantage that no tunning is necessary while it still provides excellent results [15].

Most of the above methods are in essence short-term predictors that operate on input measurements provided by road sensors or floating car data within a temporal window. The traffic flow in a time window in a specific road sensor, and the traffic flow in its neighboring sensors are two determining features to provide accurate real-time predictions [17]. They are, nevertheless, less useful for making a prediction in the long term which, as a matter of fact, cannot generally score the same accuracy level. Thus, in the event of a traffic accident, a short-term prediction model should be precise providing the vicinity measurements of traffic flow in the subsequent instants of the aforementioned accident. This is very useful for a road user, who can plan ahead his trip. On the other side, long-term predictions allow users to have a global insight of the traffic at any time, although the prediction will unlikely be accurate in the case of an atypical event (e.g. an accident). This is the reason why current research in traffic flow forecasting models is mainly focused on predicting road traffic from minutes to hours into the future. By contrast, long-term prediction models forecast traffic from several hours to days or weeks into the future, and are useful for designing public transportation policies or road planning.

In the short term, delimited circumstantial events are very relevant for the prediction models. The long term has a more seasonal nature, and small events have less relevance for the prediction as long as they do not occur regularly (for instance, closing streets for a sports event). This work delves into the study of such seasonalities and their impact on predictive models by developing the hypothesis that the average traffic flow on a certain day at an specific hour may work as a good estimator of the future traffic flow at the same hour on similar future days [18]. Therefore, the long term approach needs a characterization of the different types of days that

978-1-5090-0223-8/16/$31.00

are seasonally similar [19]. After clustering and separating the different types of days, two prediction methods – a naïve historical average and a supervised learning approach – will be shown to enhance their results when tested over real data from the city of Madrid (Spain) by virtue of the information unveiled by the cluster arrangement.

II. Data Description

Traffic prediction models usually rely on traffic variables provided by a network of road sensors. These sensors, known as Automatic Traffic Recorders (ATR), are magnetic loops embedded underground that are able to count the vehicles passing over them. It is possible to compute different metrics related to this measurement, such as speed, occupancy or traffic flow. This work focuses on traffic flow, which is the most used one when predicting road features [20].

Madrid (Spain) is one of the cities taking part in the MoveUs pilot, and has been selected for the specific task of mobility pattern analysis. The city council of Madrid has published readings of 3700 loops around the city captured every 5 minutes [21]. They also provide aggregated readings in intervals of 15 minutes with a larger historical depth. Due to the inherent search for patterns scoped by this work, we have selected the latter dataset for the sake of a larger amount of historic data, even though the granularity is lower. From all the data available we have extracted the traffic flow data, measured in vehicles per hour, from January to November 2015.

In this research we have focused in four loops placed in different areas in Madrid (Table I). Data have been grouped by loop and day, thus having 334 instances per loop, one for each day within the extracted time frame. After data cleansing (deletion of all instances with missing entries), the number of instances differs between loops. Each instance is characterized by the date and 96 entries, each for the every traffic flow readings captured on the date. The rationale for grouping the data in this way is twofold: first, loops present different behavioral patterns depending on their location (Figure 1); second, each loop will have its own prediction model.

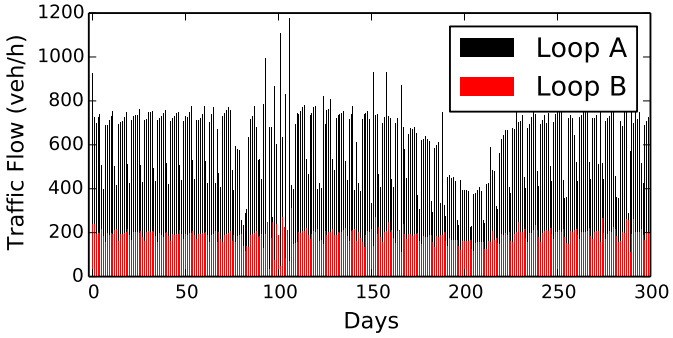

Fig. 1. Comparison between two loops of the day-average traffic flow along the year. The x axis represents each one of the days in the sample. The loop in the city center (A) has well-defined periods, while the loop in a residential area (B) is more stable through the year.

Even though other possible temporal groupings might hold, grouping by day is the way to produce understandable clusters

TABLE I
Description of the loops under consideration

Loop	Location	Details
A	Av. Pablo Iglesias and C. Juan Montalvo	City centre, intersection in a 4-lane road
B	Av. Rafael Alberti and av. de la Albufera	Urban residential area
C	Facultad Biológicas Complutense Univ.	University Campus
D	Cuesta de San Vicente Plaza España	Next to one of the main tourist attractions

composed by similar days in terms of traffic flow behavior. Furthermore, this premise can be assessed by simple visual inspection: as noted in Figure 2 there are clear differences between workdays and weekends. The question to be addressed by this work is whether this visually unveiled pattern spans beyond the mere distinction between workday and weekend.

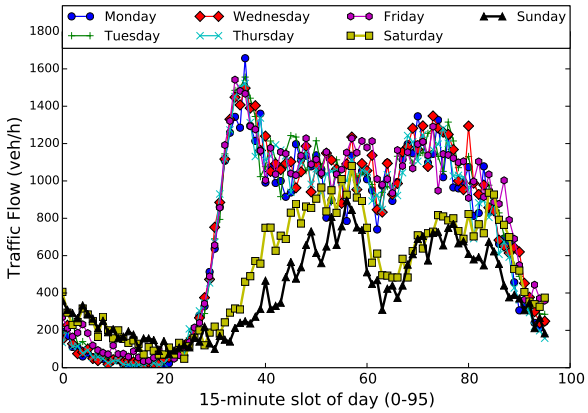

Fig. 2. Daily traffic flow per day of the week in a loop located in the city center of Madrid. Workdays tend to group, while Saturdays (yellow) and Sundays (black) follow different distributions.

Several evidences buttress the above statement: for loop A in the city center (black plots in Figure 1) there are also differences among distinct spans of the year: weeks from January to April are similar for both workdays and weekends, but August is clearly different. Furthermore, special days can be also found in mid April, coinciding with the Easter holidays. Public holidays are noticeable in other periods of the year. Loop B (red plots in the same figure) is located in a urban residential area, as evinced by a significantly lower traffic flow and less noticeable peaks, yet still featuring a remarkable decrease in weekends or public holidays. The clustering is intended to detect these typicalities that are found to vary between loops.

III. Proposed Method

The long-term prediction introduced in this manuscript is based on preprocessing the historic data into clusters that will improve the performance of the prediction algorithms. As shown in Figure 3, the process starts with a first data aggregation and preprocessing step where the information captured by the road sensor is prepared prior to clustering.

Then clusters are obtained so as to determine typical traffic patterns within the data and the factors that define such types. Finally, data are disaggregated and returned to their original shape. With this now enriched dataset, we evaluate different prediction algorithms.

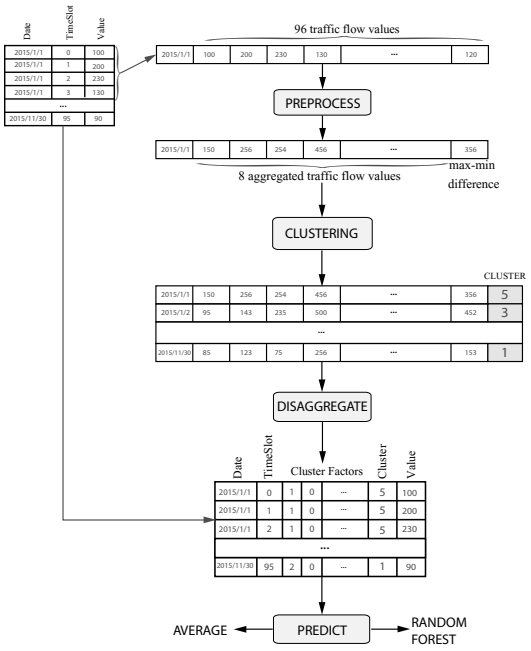

Fig. 3. Schematic diagram of the proposed method.

A. Data Preprocessing

With the grouping of traffic flow data in days we have one dataset per loop, each composed by between 250 and 330 instances depending on the result of the data cleansing. Each instance has 96 features additionally timestamped with a date. Since the high proportion of features to instances may yield model over-fitting [22], a dimensionality reduction has been performed. Days are divided into 15-minute slots, but the time series data show grouped similarities: for each loop, there is a typical day distribution of traffic flow with different parts of the day. This distribution changes in special days such as weekends, public holidays or days when great event are held. These special days are the intended results of the clustering stage. All days within the same cluster will feature a similar distribution of their traffic flow. Aggregating the 15-minute slots into n-hour slots decreases the noise produced by outliers, and depending on the size of the slots, maintains the characteristics of the typical distribution of the cluster. After trying different time aggregations, 15-minute slots are aggregated into 3-hour slots. Besides this aggregation we have appended the difference between the maximum and minimum daily traffic flow value as a new feature. This will contribute to an easier discrimination of days with greater peaks – usually working days – from the rest.

B. Clustering

The historic average can be conceived as the straightforward approach for traffic flow forecasting, although it is necessary to define which values intervene in the computation of the average. It is clear that if all historic values participate in this computation, the prediction will be far from accurate if data are not statistically stationary enough. A more refined way to estimate the average is by slot of time. In our dataset this would imply calculating the average of e.g. all the first 3-hour slots of each day, which would constitute the predicted value for any traffic flow happening in the first three hours of any day. An obvious error stems from this approach: it is very unlikely that the first 3 hours of a Sunday in August undergo a similar traffic flow to the first 3 hours of a Wednesday in March. The day of the week, the month, the holidays or the event calendar are factors that underline different day patterns. In other words, clusters that will help the average or any other prediction method perform efficiently. These factors are different for each loop, as seen in Figure 1.

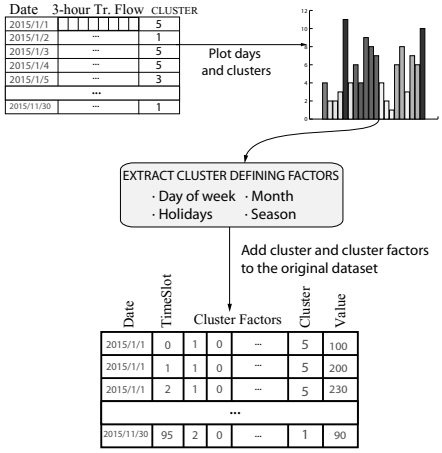

Fig. 4. Reconfiguration of the dataset after clustering. The detected cluster factors are the day of the week, month, season of the year and holidays

We have automated the process of discovering their relevance by using a density-based clustering algorithm, – specifically, DBSCAN [23] – which classifies each day in groups or clusters based on a metric of similarity (in our case, Euclidean distance). The optimal number of clusters is defined following the so-called Elbow approach [19]. The model needs to set the maximum distance between instances for them to be considered as part of the same cluster. As the ranges of traffic flow values differ widely from one loop to another, this parameter is adjusted manually following the aforementioned steps, resulting in 5 clusters plus one noise cluster that groups the days not fitting in any other cluster. Once clusters have been found, underlying factors that are common to their constituent instances or days (*commonalities*) are explored. Such factors will be added to the original dataset, that will have then as features the date, the time of the day, the value of traffic flow, and a series of features defined by the cluster factors (Figure 4). These extended dataset will be the basis for subsequently performing predictions.

C. Prediction and evaluation

The prediction of future values will be made for individual timestamps, i.e. the traffic flow of April 25th 2016 at 10.05 am. The day-based data organization used for clustering is not useful on this purpose, hence the data will be reverted again to their original matrix arrangement by adding the informative features produced by the previous clustering stage.

As aforementioned in the introduction, two different prediction models are considered: a naïve historical averaging method and a supervised learning model. Regarding the latter, Random Forest Regressors [24] have been selected, which builds upon an ensemble of regression trees whose predicted values after bagging and training are aggregated via averaging. Random Forests have been theoretically proven to reduce the variance of the prediction and avoid over-fitting at the cost of a slight penalty in the bias of the overall predictor [25]. We will evaluate the results between of these approaches with and without cluster information. To this end, a leave-one-out cross-validation strategy will be adopted: the prediction is made individually for the 96 15-minute-time-slot traffic flow values of one single day, training with the values of the rest of the days. As such, the prediction $\widehat{F}_d\,(n)$ for the left-out day d, time slot n and the historical averaging method is the average of all traffic flow measurements sharing the same cluster features and slot of day in the training set, i.e.

$$\widehat{F}_d\,(n) = \frac{1}{|\mathcal{C}(d)| - 1} \sum_{\substack{k \in \mathcal{C}(d) \\ k \neq d}} F_k\,(n) \quad \forall d \in \mathcal{C}(d), \qquad (1)$$

where $|\cdot|$ stands for cardinality and $\mathcal{C}(d)$ denotes the set of days within the dataset that share the same values for the commonalities or factors derived from the clustering space. As for the Random Forest Regressor, an individual model is trained and evaluated for each slot of the day.

IV. Experimental Results

A. Clustering Results

We begin the analysis of the experiments performed over the data captured in the city of Madrid (Spain) by first addressing Figure 5, where the clustering results for the considered loops are depicted. At this point it is important to recall that the targeted goal of the clustering approach is to detect commonalities within the instances compounding each of the clusters found in the process. In Figure 5 (above) it is possible to observe the differences among the different types of days for loop A, located in the city center.

The clustering stage applied to the rest of loops renders analogous results with the same number of clusters and similar distribution of the days within them, as seen in the rest of subplots in Figure 5. Although in some loops the order of the clusters is reversed (as in loop C, third subplot), it is important to remark that they are formed by the same types of days. However, other loops behave quite differently. Loop D (fourth subplot in Figure 5) exemplifies this observed pattern: it is located next to an important tourist attraction, and the differences between Saturdays and Sundays are hardly noticeable. Furthermore, during holiday periods such as Easter, loop D undergoes a traffic flow similar to workdays, although it is detected as *noise*[1] by the clustering algorithm. The results shown in Table II evince the similarity among the clusters in different loops around the city, with the exception of loop D due to the already commented location particularity.

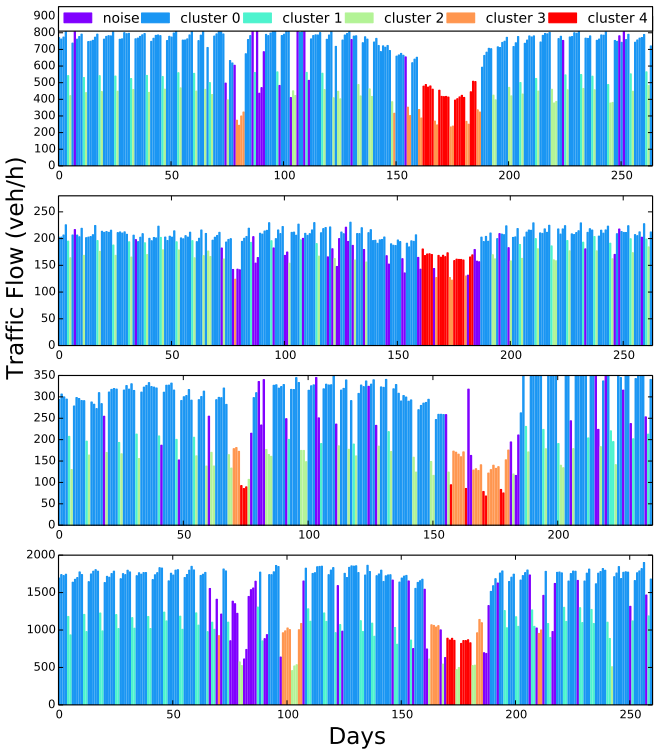

Fig. 5. From top to bottom, clustering results for loops A, B, C and D. The x axis represents each day of the sample, and the different clusters are represented by colors. Although the average traffic flow values are lower in loop B than in loop A, the detected clusters are similar. The detected clusters are similar in cluster C, but the clusters for week days and weekends of august are reversed. Clustering the days for the loop D, next to a tourist attraction. Increased noisy instances, and different holiday periods than in the other loops.

TABLE II
Number of days in each cluster per loop

Cluster Identifier	Loop A	Loop B	Loop C	Loop D
noise	18	38	27	38
0	150	152	128	130
1	22	21	19	55
2	39	30	36	12
3	17	5	19	16
4	18	17	9	9

By analyzing the instances (days) included in each discovered cluster it is possible to infer the factors that define whether a day is part of a cluster without resorting to its corresponding traffic flow measurements. Cluster 0 is the one corresponding to workdays, and tends to be stable in all loops from January to April. Saturdays and Sundays are separated

[1]Density-based clustering algorithms declare a given sample as *noise* when the number of other instances at distance less than ϵ is below a specified minimum neighbor size, equal to 5 in the performed experiments.

into two clusters, 1 and 2. Interestingly yet expectedly, some local holidays happening in Monday are also clusterized as Sunday. For loop D, there is no difference between Saturdays and Sundays. Clusters 3 and 4 are reversed in loop C, but contain the same types of days than in other loops: workdays of August, for cluster 4 (3 in C), and Easter, August weekends, and public holidays in July for cluster 3 (4 in C). Easter is detected as noise in D: in this case the traffic is not similar to the one occurring in public summer holidays, hence it cannot constitute a cluster due to its low cardinality. Figure 5 (bottom) shows the traffic flow in this location during Easter is much closer to a working day traffic flow than in the other loops. The differences between clusters in different loops are not relevant since each loop will have its own prediction model. Based on this observation, the most relevant factors derived from the clustering process are:

- Day of the week: the day of the week will be divided into working day (0), Saturday (1) and Sunday (2).
- Month: the month appears to be relevant and similar for all the loops: August is different than the rest of the year, and the three first months are stable. The tag will contain the month number.
- Public holidays: the public holidays make the difference. Easter and days like May 1st, July 25th or October 12th are different than the rest of the days for all of the loops, except when they are in a weekend. Local holidays such as June 4th are less relevant, except if they produce long weekends. Public holidays will be another feature for the prediction model.

Each independent value of the original sample has been tagged with these factors as an input to the prediction model. Some of the observed factors are combinations of these three. Weekends of August conform a cluster along with other public holidays. We have not created features for them, expecting the model to be able to detect these special days. There are other factors that are particular to each loop, such as the presence of traffic works or important events. We expect our model to be generalizable for all loops, and consequently these other factors have not been taken into account.

B. Prediction Results

The naïve historic average approach is made for each cluster and time slot of day. The average of all the traffic flow measurements of each time slot in a cluster is obtained and used as the prediction. The noise cluster average is also used as a reference for the noisy instances. The error will be higher in these instances, but the machine learning model can obtain a better performance. This machine learning model is based in random forest algorithm. Like with the average, a submodel has been trained and tested for each slot of day. The standard error of the estimates of these two models averaged over slots and compared to their non-clustered versions are shown in Table III, which should be understood jointly with the average flow values recorded by each loop: 606.8 (A), 188 (B), 244.2 (C) and 1363 (D) vehicles per hour.

TABLE III
STANDARD ERROR OF THE PREDICTIONS (IN VEHICLES/HOUR, AVERAGED OVER SLOTS) WITH AND WITHOUT CLUSTER-DERIVED INFORMATION

	σ_{est}	**Loop A**	**Loop B**	**Loop C**	**Loop D**
Without clusters	Average	238.89	51.92	112.03	568.05
	Random Forest	239.75	52.11	112.49	570.02
With clusters	Average	105.74	38.02	57.31	264.25
	Random Forest	102.85	37.01	58.20	285.31

After clustering, the error of both predictions is reduced to values close to each other. The errors are proportional to the average traffic flow of each loop, but in all four cases they are around the half of the predictions without clusters. These results confirm the relevance of the clustering as a pre-processing for the prediction model. Contrary to expectations, the model has not improved the results of the historic average in a significant way. With or without clusters, both approaches produce similar predictions. The historic average model is improved by feeding it with clustering information; with the same information, the random forest model outperforms the original random forest model, but not the improved average model. Figure 6 shows the overlapping between the two predictions per slot, with the mean prediction of each model and the 80 % percentile for each slot.

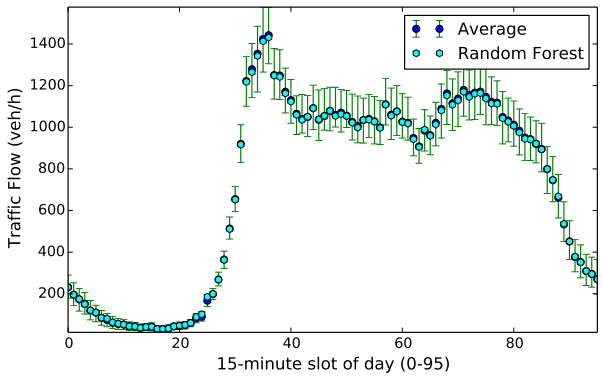

Fig. 6. Comparison of historic average and random forest predictors in cluster 0. Average (dark) and random forest (light) predictions overlap for most of the values: the outputs of both models are very similar. Vertical error lines delimit the 80% percentile of the predicted values for each slot.

The prediction results have been plotted for an individual day of each cluster in Figure 7. The lines corresponding to the models with cluster information are fitted to the real values lines. The days in the noise cluster behave differently, as they are a varied range of days with very different traffic flow distributions. It is in these days where errors become more noticeable.

V. CONCLUSIONS AND FUTURE WORK

The traffic flow in different measuring points of a road network can be characterized via clustering. This characterization allows performing long-term predictions that are not as accurate as short-term forecasts, but that can help decision makers taking decisions for a more efficient traffic management. The results obtained in this work empirically

show the relevance of the clustering for a long-term prediction model, and the good performance of a naïve average approach when the dataset is preprocessed conveniently. There is though room for improving the results by refining the clustering stage.

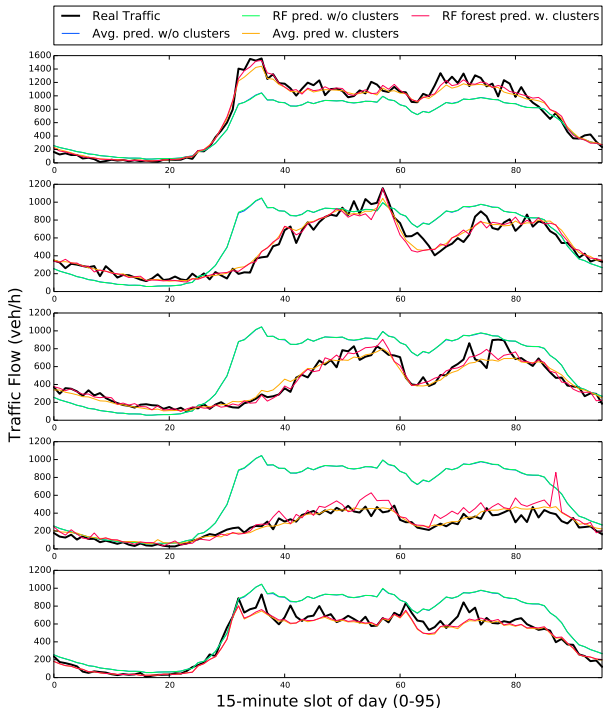

Fig. 7. Comparison of average and random forest predictions with and without cluster-derived information in a day pertaining to each cluster: workdays (subplot 1), Saturdays (subplot 2), Sundays (subplot 3), public holidays and August weekends (subplot 4) and August workdays (subplot 5).

The main risk of basing a prediction model in historic data lies on its potential misadjustment to circumstantial events or characteristics of the road. Including periodic events in the dataset is an obvious next step to improve the prediction model. Sports events, demonstrations or parades usually imply the closing of roads, and they have a relevant impact in the traffic surrounding the location of the event. However, in large cities like Madrid they rarely affect the traffic of the entire urban road network. While the clustering factors that we have used in this paper are general for any location in the city (e.g. it is Saturday everywhere), handling the events require location sensitiveness. Another relevant improvement for the model is ageing the data, granting less relevance to old data for the model training. This helps when there are road works, making the traffic denser, or a new road is built, making the traffic of the old road less dense. Although traffic is clearly seasonal, it changes in long-term periods. Ageing the data makes the model more context-aware. Big Data capabilities will be also explored for extending this study to a higher number of loops.

ACKNOWLEDGMENT

The research leading to these results has received funding from the European Communitys Seventh Framework Programme (FP7/2007-2013) under Grant Agreement n. 608885.

REFERENCES

[1] S. P. Hoogendoorn, P. H. Bovy, "State-of-the-art of Vehicular Traffic Flow Modeling", Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering, Vol. 215(4), pp. 283–303, 2001.

[2] M.S. Ahmed, A. R. Cook, "Analysis of Freeway Traffic Time-Series Data by using Box-Jenkins Techniques", *Transportation Research Record*, N. 722, pp. 1–9, 1979.

[3] B. M. Williams, L. A. Hoel, "Modelling and Forecasting Vehicular Traffic Flow as a Seasonal ARIMA Process: Theoretical Basis and Empirical Results", *Journal of Transportation Engineering*, Vol. 129, N. 6, pp. 664–672, 2003.

[4] S. Vasantha Kumar, L. Vanajakshi, "Short-Term Traffic Prediction using Seasonal ARIMA Model with Limited Input Data", *European Transport Research Review*, pp. 7-21, 2015.

[5] M. VanderVoort, M. Dougherty, S. Watson, "Combining Kohonen Maps with ARIMA Time Series Models to Forecast Traffic Flow", *Transportation Research Part C: Emerging Technologies*, Vol. 4, pp. 307-318, 1996.

[6] W. Hu, Y. Liu, L. Li, "The Short-Term Traffic Flow Prediction based on Neural Network", *International Conference on Future Computer and Communication*, Vol. 1, pp. 293–296, 2010.

[7] G. Tan, H. Shi, F. Wang, C. Deng, "Short-Term Traffic Flow Prediction based on Parallel Quasi-Newton Neural Network", *International Conference on Measuring Technology and Mechatronics Automation*, pp. 305–308, 2009.

[8] K. Y. Chan, T.S. Dillon, J. Singh, E. Chang, "Neural-Network-based Models for Short-Term Traffic Flow Forecasting using a Hybrid Exponential Smoothing and Levenberg-Marquardt Algorithm", *IEEE Transactions on Intelligent Transportation Systems*, Vol. 13, N. 2, pp. 644–654, 2012.

[9] H. Dia, "An Object-oriented Neural Network Approach to Short-Term Traffic Forecasting", *European Journal of Operational Research*, Vol. 131, N. 2, pp. 253–261, 2001.

[10] K. Kumar, M. Parida, V. K. Katiyar, "Short Term Traffic Flow Prediction for a Non Urban Highway using Artificial Neural Networks", *Procedia - Social and Behavioral Sciences* vol. 104, pp. 755–764, 2013.

[11] H. Su, L. Zhang, S. Yu, "Short-Term Traffic Flow Prediction based on Incremental Support Vector Regression", *International Conference on Natural Computation*, pp. 640–645, 2007.

[12] D. Zeng, J. Xu, J. Gu, L. Liu, G. Xu, "Short-Term Traffic Flow Prediction based on Online Learning SVR", *Workshop on Power Electronics and Intelligent Transportation System*, pp. 616–620, 2008.

[13] D. Zeng, J. Xu, J. Gu, L. Liu, G. Xu, "Short Term Traffic Flow Prediction using Hybrid ARIMA and ANN Models", *Workshop on Power Electronics and Intelligent Transportation System*, pp. 621-625, 2008.

[14] X. Guo, F. Deng, "Short-Term Prediction of Intelligent Traffic Flow on BP Neural Network and ARIMA Model", *International Conference on E-Product E-Service and E-Entertainment*, pp. 1–4, 2010.

[15] H. R. Kirby, S. M. Watson, M. S. Dougherty, "Should We Use Neural Networks or Statistical Models for Short-Term Motorway Traffic Forecasting?", *International Journal of Forecasting*, Vol. 13, pp. 43–50, 1997.

[16] C. P. Van Hinsbergen, J. W. Van Lint, F. M. Sanders, "Short Term Traffic Prediction Models", *ITS World Congress*, Beijing, China, 2007.

[17] W. Min, L. Wynter, Y. Amemiya, "A Road Traffic Prediction with Spatio-Temporal Correlations", IBM Research Division, 2007.

[18] E. Chung, N. Rosalion, "Short Term Traffic Flow Prediction", *Australian Transportation Research Forum*, Hobart, Tasmania, 2001.

[19] W. Weijermars, E. van Berkum, "Analyzing Highway Flow Patterns using Cluster Analysis", *International IEEE Conference on Intelligent Transportation Systems*, pp. 831–836, 2005.

[20] L. Yisheng, Y. Duan, W. Kang, "Traffic Flow Prediction with Big Data: a Deep Learning Approach", *IEEE Transactions on Intelligent Transportation Systems*, Vol. 16, N. 2, pp. 865–873, 2015.

[21] Madrid Open Data portal, http://datos.madrid.es, accessed on Jan. 2016.

[22] P. Mitra, C. A. Murthy, S. K. Pal, "Unsupervised Feature Selection using Feature Similarity", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 24, N. 4, pp. 1–13, 2002.

[23] M. Ester, H. Kriegel, J. Sander, X. Xu, "A Density-based Algorithm for Discovering Clusters", *International Conference on Knowledge Discovery and Data Mining*, pp. 226-231, 1996.

[24] L. Breiman, "Random Forests", *Machine Learning*, Vol. 45, pp. 5–32, 2001.

[25] G. Biau, "Analysis of a Random Forests Model", *Journal of Machine Learning Research*, Vol. 13, N. 1, pp. 1063-1095, 2012.