sciendo

Transport and Telecommunication, 2020, volume 21, no. 2, 119–124
Transport and Telecommunication Institute, Lomonosova 1, Riga, LV-1019, Latvia
DOI 10.2478/ttj-2020-0009

A SCALABLE APPROACH FOR SHORT-TERM PREDICTIONS OF LINK TRAFFIC FLOW BY ONLINE ASSOCIATION OF CLUSTERING PROFILES

Alessandro Attanasi[1], Marco Pezzulla[2], Luca Simi[3], Lorenzo Meschini[4], Guido Gentile[5]

[1]PTV Group SISTeMA
Rome, Italy, via Bonghi 11b
alessandro.attanasi@ptvgroup.com
[2]PTV Group SISTeMA
Rome, Italy, via Bonghi 11b
marco.pezzulla@ptvgroup.com
[3]PTV Group SISTeMA
Rome, Italy, via Bonghi 11b
luca.simi@ptvgroup.com
[4]PTV Group SISTeMA
Rome, Italy, via Bonghi 11b
lorenzo.meschini@ptvgroup.com
[5]Sapienza Università di Roma, DICEA
Rome, Italy, via Eudossiana 18
guido.gentile@uniroma1.it

Short-term prediction of traffic flows is an important topic for any traffic management control room. The large availability of real-time data raises not only the expectations for high accuracy of the forecast methodology, but also the requirements for fast computing performances. The proposed approach is based on a real-time association of the latest data received from a sensor to the representative daily profile of one among the clusters that are built offline based on an historical data set using Affinity Propagation algorithm. High scalability is achieved ignoring spatial correlations among different sensors, and for each of them an independent model is built-up. Therefore, each sensor has its own clusters of profiles with their representatives; during the short-term forecast operation the most similar representative is selected by looking at the last data received in a specified time window and the proposed forecast corresponds to the values of the cluster representative.

Keywords: forecast, clustering, Big Data, scalable architecture

1. Introduction

The amount of mobility data available in modern cities is growing exponentially and consequently the need of properly collecting and organizing them is crucial for any forthcoming analysis and usages. The Big Data paradigm (McAfee, 2012) is taking place in urban contexts around mobility data as indeed described by the five Vs: Volume, Velocity, Variety, Veracity and Value. Volume refers to the vast amounts of data generated constantly that prevent the usage of traditional database technology for storage and analysis. Velocity refers to the speed at which new data is generated, for example by mobile phones or GPS probes. Variety refers to the different type of data we can use, like traffic counts, speed data, GPS data, mobile data, pictures and video streams from traffic light camera. Veracity refers to the wide spectrum of quality and accuracy of the data acquired. And finally, Value is the capacity to extract from this huge amount of data some gain without which any big data initiatives will have an unclear understanding of costs and benefits.

In this framework, modern traffic management control room needs to solve among several transportation challenges the one for providing real-time short-term forecast in order to enable decision support systems and to mitigate traffic congestion. For that reason, having high accuracy forecast methodologies is mandatory, dislike of the big amount of data available in real-time that requires fast

computational capabilities. To properly tackle these big data, modern IT infrastructure must be designed in a way that they can be scaled in proportion to the volume and velocity of the input data. This scaling is achieved horizontally, in contrast to vertical scaling; indeed, increasing the hardware capabilities of a single computational server as done for vertical scaling is totally useless when dealing with big data because there is not any single hardware capable to process all the data together due to both memory and computing constraints. The processes need to be parallelized on several worker stations, and this requirement has an impact on the software architecture and on the algorithms themselves.

The main contribution of this research is to propose and test on real life data a real-time short-term forecast methodology considering as input flow data coming from loop detectors. The proposed methodology is conceived to be naturally horizontally scalable and despite its theoretical simplicity the forecast accuracy is anyhow good and comparable with the ones achieved by more complex approaches (Gentile, 2011), (Attanasi, 2017).

The rest of the paper is structured as follows. Section 2 describes the forecast methodology and results achieved. Section 3 shows in detail the application of the proposed methodology to a real-life data set. Section 4 results and proposes further improvement directions.

2. Methods and Results

Unlike network transportation models, which derive variables evolution by physical laws regarding the propagation of flows and the behaviour of users (Gentile, 2011), the proposed method exploits only the observed data trend to provide the predictions. The only analyzed quantity is the traffic flow, but the methodology can be generalized to other traffic observables.

A key ingredient of the proposed methodology is the fact that a loop detector is considered as a standalone device, so its embedding into the transportation graph and its relationship with other devices due to the correlations of traffic information are intentionally disregarded. The driver of this decision is the capability to easily scale the problem parallelizing the computation. Indeed, considering each loop detector as a single entity totally uncorrelated from other ones, it is straightforward to achieve high performance by simply having in parallel several worker stations capable to process only few detectors each.

The methodology has been applied to a study area in the city center of Turin, where 658 count locations cover the main zones of the street network. Each count location is geo-referenced to a directed arc of the network, and the flow counted by each of them is only the flow travelling in one direction of the street. The dataset is constituted by two years of flow data, 2015-2016. The frequency rate is 1 measure every 15 minutes, and the dataset contains missing and meaningless values too. Data of the year 2015 has been used for offline clustering; and data of the year 2016 are used to mimic real-time behavior and to assess the forecast quality.

2.1. Offline Clustering

The analysis of flow data is a common task in transportation analysis for achieving several purposes such as proper demand matrix calibration, identification of specific space and time bottlenecks into the system, understanding common congestion patterns in both space and time again. Capturing the same average behavior for a set of days is a key aspect, and indeed for those aggregation of similar days the concept of day-type is introduced exactly meaning that all the single calendar days belonging to the same day-type behave in the same average way. For example, a common day-type grouping is the following: Working-Days, Pre-Holidays, Holidays, and divided between the summer and the winter period. As can be immediately recognized, this naïve grouping can be wrong a priori because there is no guarantee that this splitting ensures the similarity among profiles of the same day-type. Because the forecast methodology that we propose is deeply dependent from this grouping, we must find a way that can be as much as possible accurate in this clustering operation. And because of our target is to have a dedicated forecast for each loop detector without any relationship with other loop detectors, the approach is to perform several clustering procedures one for each detector. Clearly in that way the abstract concept of day-type applicable to the whole network is not valid anymore, but it does not hurt the forecast procedure. We will have several clusters for each detector, and they can be different in term on numerosity and composition, but this is only a benefit for our methodology because in this way the high specialization of each detector will provide a better local forecast despite using an overall average clustering. It is indeed highly common to observe on different loop detectors a not equal traffic pattern.

The clustering algorithm we selected for this research is the Affinity Propagation (Dueck, 2008a), (Dueck, 2007a), one of the *de facto* standard clustering nowadays widely used in several applications (Dueck, 2008b), (Dueck, 2007b), (Lazic, 2009). The two main reasons for which we selected it instead of others (like k-means (David, 2007), DBSCAN (Ester, 1996), Spectral Graphs (Chun, 1997)) is the fact that it is an unsupervised technique where the number of clusters is not specified by the user, and because it has really few parameters to be fine-tuned, mainly one, and this helps a lot for the setup of a new system reducing its calibration.

The clustering operation is performed on daily profiles of flow data. These profiles span the whole day from 00:00 to 24:00 with data at most every 15 minutes. A crucial point of the clustering procedure is the selection of an appropriate and meaningful similarity function between two profiles. We adopted the normalized correlation; if p_1 and p_2 are two different profiles, the similarity function is defined as:

$$s(p_1, p_2) = p_1 p_2 / (\| p_1 \|_2 \| p_2 \|_2), \quad (1)$$

where $p_1\ p_2$ is the dot product between p_1 and p_2 and $\| \|_2$ is the Euclidean norm.

Because the Affinity Propagation belongs to the family of medoids clustering (Madhulatha, 2012), it will return for each cluster the exemplar daily profile, i.e. the most representative daily profile.

It is worth to observe that the similarity function as defined in (1) will drive the Affinity Propagation in creating clusters where the profiles of the same cluster have roughly the same shape: this means that if into the dataset there are profiles with same shapes but with a wide range of values (because for example the volume of people travelling on the same detector varies a lot from summer to winter) they will be grouped together. This detail is important for the online forecast as we will see in section B.

The clustering of the data of each detector can run in parallel to all the other ones, so at the infrastructure level it is easy to scale horizontally this offline computation simply by having multiple worker computing engines. And if N is the average number of daily profiles of a loop detector, and assuming several available computing resources equal to the number of detectors, the running time of the process scales quadratically with N, simply because you need $O(N^2)$ time for elaborating the similarity matrix and $O(N^2)$ time for executing the Affinity Propagation algorithm.

2.2. Online Forecast

Once the offline clustering process is done, we can start providing an online short-term forecast. The way we devised it is again fully parallelizable with respect to the detectors because to get the forecast of each detector the methodology needs only data of the corresponding detector regardless what happen to the others. This behavior allows a high horizontal scalability for the short-term forecast, and this capability is of great importance in modern cities where the amount of real time data is so huge that fast algorithms are required.

The online forecast computation for a given detector is triggered once a new data is received; at this point in time the algorithm retrieves a configurable window of recent past data (usually one hour backward) and by using the same similarity function by which the offline clustering was performed, the most similar cluster exemplar to these time window data is selected. And as forecast result for the desired forecast horizon (usually the next hour), the corresponding interval of the profile of the cluster representative is provided. But as we said in section *A* the similarity function we use matches two profiles based on their shapes, so it can happen that real-time data despite having the same shape of the exemplar, they can be much different in term of absolute numbers. To cope with this situation, the representative profile is shifted toward the real-time data of an amount equal to the difference between it and the real-time data at the time instant of the computation triggering.

It is worth to note that the online association is recomputed every time new data arrive, so this flexibility will enhance the result because even if the cluster exemplars represent average traffic profiles, the continuous switching capability will deal better with the daily traffic fluctuation.

2.3. Forecast Quality Indicators

In order to judge the quality of the forecast provided by the proposed methodology it is also important to clearly state the used Key Performance Indicator (KPI). Despite standard indicators adopted for similar purposes in the literature, like MAE (Mean Absolute Error), or MAPE (Mean Absolute Percentage Error), or RMSE (Root Mean Square Error), we focus only on a specific transport engineering KPI known as GEH by the name of its creator Geoffrey E. Havers (Transport for London, 2016). The GEH formula is given by:

$$GEH = \sqrt{\frac{2(M-F)^2}{M+F}}, \quad (2)$$

where M is the measured traffic volume, while F is the forecast traffic volume. Note that both measurements must be expressed as vehicles per hour. Although the GEH seems to be a chi squared test, it is not a true statistic indicator, but it is an empirical formula demonstrated to be useful in several traffic situations. Because in real world transportation systems the traffic volumes on a given count section can vary widely, it is impossible to set a single threshold value to judge the percentage variation of the volumes. But being the GEH a not linear formula, a single acceptance threshold on GEH can be used over a wide range of traffic volumes.

The GEH formula is often used for evaluation of transportation models to evaluate the achieved level of calibration of an offline model, typically built for planning purposes. This means that the average of measured volumes on a given count site and on a specific day-type are compared against the simulation volumes produced by the model usually by running a static traffic assignment. In this setting the acceptance criteria, recognized valid though a large amount of experimental data, are: at least 85% of the volumes predicted by the model should have a GEH less than 5.0; GEHs between 5.0 and 10.0 should be investigated, while values greater than 10.0 indicate a serious problem into the model that with high probability has to be recalibrated (Transport for London, 2016).

But our forecast methodology is dynamic because it changes as a function of the time and it produces more forecast results for several horizons. For that reason, we must generalize (2) to take this information into account:

$$GEH(t;h) = \sqrt{\frac{2(M(t)-F(t-h))^2}{M(t)+F(t-h)}}, \quad (3)$$

where t is a specific time instant of the day, and h is the forecast horizon. Despite the simplicity of this generalization, to the best of our knowledge there are not common accepted criteria for evaluating a dynamic forecast. What is a priori expected is that the bigger the horizon h the worst the GEH is, and for peak hour traffic congestion time the GEH must be higher; all of that because the forecast for bigger horizons and for heavily congested time are expected to be much more difficult. Within this paper we would like to provide a guideline for setting in a meaningful way the acceptance criteria in the dynamic setting, and for doing this it is crucial to analyze the naïve forecast and its performance.

The path we pursued for defining these criteria are methodologically the following: given several days in a year and many flow daily profiles on several detectors located in an urban context, we grouped the data by detector. For each detector we computed the GEH at a given forecast horizon and for each time instant of the daily profile the average GEH is elaborated. As a final aggregation step, we considered the mean of the GEH of each detector and we recomputed some descriptive statistics (mean, standard deviation, percentiles) reducing over the detector dimension. In this way we can state what is the GEH threshold daily profile for a fixed horizon once we pick the percentiles of this last statistics. For example, if we chose 90 as percentiles it means we are asking to have at least 90% of detectors with an average GEH under the threshold value. In section 3 detailed results as shown considering real data.

3. Numerical Experiments

The described methodology has been applied to a real study case, where 658 count locations cover the main zones of the street network of the city of Turin and its surroundings. The dataset is constituted by two years of flow data, 2015-2016. The frequency rate is 1 measure every 15 minutes, and the dataset contains missing and meaningless values too. Data of the year 2015 has been used for offline clustering; and a subset of the data of the year 2016 is used to mimic real-time behavior and to assess the forecast quality.

Main properties of the data set analyzed for the assessment are summarized in Table 1.

Table 1. Raw Data Properties

Distinct days	100
Total number of loop detectors	658
Average number of days per detectors	70
Average percentage of data per detector	74%

First, we analysed the forecast quality indicators for the naïve forecast methodology, and then we did the same for our methodology. The naïve approach consists in a simple flat repetition of the current data for the next forecast horizons.

In Figure 1 are shown for each forecast horizon and time of the day the 90 percentiles of all mean GEH values of all detectors. Looking at those aggregated results we can state for the forecast horizon of 15 minutes that the naïve methodology is better of the proposed methodology independently of the time of the day, but it is worth to note that in both cases the average daily GEH threshold is around 4, so really good having in mind the threshold used for static simulations. For both approaches the longer the forecast horizon the worst the GEH values are, as it was expected. We can note also that the degradation of the forecast capability (intended as a worsening of the GEH value at a given time), increasing the forecast horizon is faster for the naïve methodology compared to the proposed one. Finally, our methodology is capable to handle much better peak hours congestions keeping the GEH under 9 also at one hour forecast horizon.

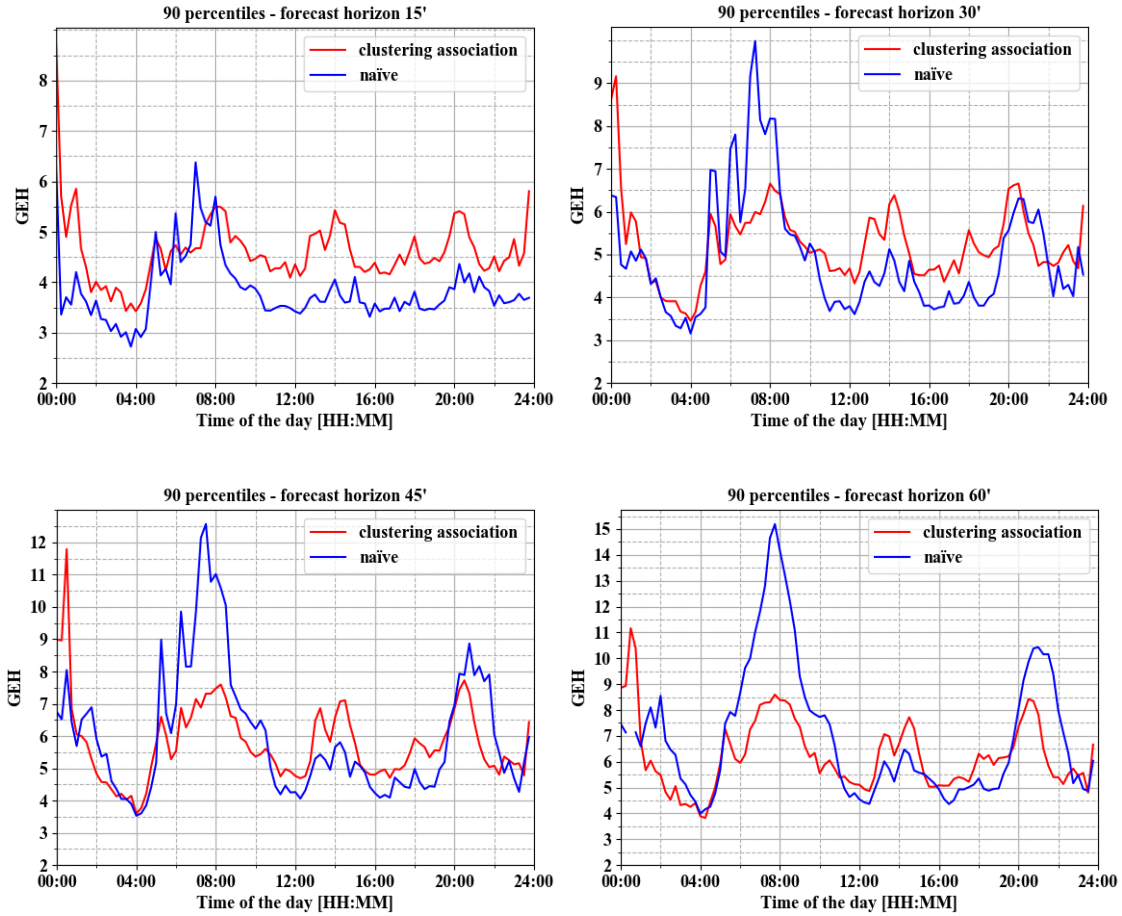

Figure 1. The figure shows for four forecast horizons (15, 30, 45, 60 minutes, respectively sub-figures a, b, c, d) the 90 percentiles daily profiles of the average GEH values of all detectors for both the naïve forecast methodology and the cluster association one

4. Conclusions

The paper addressed comparison of a short-term forecast methodology for traffic flow data based on a clustering technique against a naïve methodology. This type of comparison was done looking at GEH as indicators for the goodness of the forecast, and it was useful also for setting, by the naïve approach, some thresholds on the GEH as a function of the time to be used as reference points for whatever else forecast methodology.

Those thresholds are reported having analyzed 100 days of flow profiles on 658 detectors in an urban context, and they refer to a situation for which 90% of the detectors have an average GEH value equal to the threshold reported at a given time of the day. What is worth also to underline is that these

thresholds at a fixed time increase for longer forecast horizon, and in general are worst in correspondence of peak hour congestions. Considering the proposed methodology, we can state that it performs good for longer forecast horizons, and especially it is much more robust and stable of the naïve approach during all the time of the day and so specifically during congested peak hours.

As future directions of improvements we can see a deeper investigation of the results aggregated by day-types, a sensitivity analysis of the proposed forecast methodology with respect its parameters like the one driving the Affinity Propagation offline clustering, or the time window width of real-time data used for online forecast. And it will be of great interest to analyze how the forecast methodology performs in case of speed data.

Acknowledgements

Authors thank 5T company for sharing their raw data allowing this research.

References

1. Attanasi, A. et al. (2017) A hybrid method for real-time short-term predictions of traffic flows in urban areas, *IEEE International Conference on MODELS and Technologies for Intelligent Transportation Systems IEEE*, pp. 878-883, 2017
2. Chun, F. (1997) Spectral graph theory, *CBMS Regional Conference Series in Mathematics, 92*. Conference Board of the Mathematical Sciences, Washington. 1997.
3. David, A. and Vassilvitskii, S. (2007) k-means++: The advantages of careful seeding. *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, Society for Industrial and Applied Mathematics*.
4. Dueck, D., Frey, B. J., Jojic, N., Jojic, V., Giaever, G., Emili, A., Musso, G., Hegele, R. (2008a) Constructing treatment portfolios using affinity propagation. In: *Vingron, M., Wong, L. (eds.) RECOMB 2008. LNCS*, 4955, pp. 360–371. Springer, Heidelberg (2008).
5. Dueck, D., Frey, B. J. (2007a) Non-metric affinity propagation for unsupervised image categorization. In: *Proceedings of the IEEE 11th International Conference on Computer Vision (ICCV '11)*, October 2007, Rio de Janeiro, Brazil. IEEE Press, pp. 1–8.
6. Dueck, D. et al. (2008b) Constructing Treatment Portfolios Using Affinity Propagation, *International Conference on Research in Computational Molecular Biology (RECOMB)*, March 2008, Singapore.
7. Dueck, D. and Frey, B. J. (2007b) Non-metric affinity propagation for unsupervised image categorization, *International Conference on Computer Vision (ICCV)*, October 2007, Rio de Janeiro, Brazil
8. Ester, M., Kriegel, H. P., Sander, J. and Xu, X. (1996) A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In: *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, Portland, OR, AAAI Press, pp. 226-231.
9. Gentile, G., Meschini, L. (2011) Using dynamic assignment models for real-time traffic forecast on large urban networks. In: *Proceedings of the 2nd International Conference on Models and Technologies for Intelligent Transportation Systems*, 2011, Leuven, Belgium.
10. Lazic, N., Givoni, I. E., Aarabi, P. and Frey, B. J. (2009) FLoSS: Facility Location for Subspace Segmentation, *12th International Conference on Computer Vision (ICCV)*, October 2009, Kyoto, Japan.
11. McAfee, A., Brynjolfsson, E. (2012) Big data: The management revolution, *Harvard Business Review*, 90(10), pp. 60-68.
12. Soni Madhulatha, T. (2012) An Overview on Clustering Methods, *IOSR Journal of Engineering*, 2(4), 2012, April, pp. 719-725.
13. Transport for London (2016). *Traffic Modelling Guidelines Version 3.0*, http://content.tfl.gov.uk/traffic-modelling-guidelines.pdf, Retrieved 10-March-2016