Australasian Transport Research Forum 2017 Proceedings
27 – 29 November 2017, Auckland, New Zealand
Publication website: http://www.atrf.info

Improving adaptation and interpretability of a short-term traffic forecasting system

Rafael Mena-Yedra[1,2], Jordi Casas[1], Tamara Djukic[1], Ricard Gavaldà[2]

[1]TSS-Transport Simulation Systems, Ronda de la Universitat, 22b, 08007 Barcelona, Spain

[2]Universitat Politècnica de Catalunya, Campus Nord, C/ Jordi Girona, 1-3, 08034 Barcelona, Spain

Email for correspondence: rafael.mena@aimsun.com

Abstract

Traffic management is being more important than ever, especially in overcrowded big cities with over-pollution problems and with new unprecedented mobility changes. In this scenario, road-traffic prediction plays a key role within Intelligent Transportation Systems, allowing traffic managers to be able to anticipate and take the proper decisions. This paper aims to analyse the situation in a commercial real-time prediction system with its current problems and limitations. The analysis unveils the trade-off between simple parsimonious models and more complex models. Finally, we propose an enriched machine learning framework, Adarules, for the traffic prediction in real-time facing the problem as continuously incoming data streams with all the commonly occurring problems in such volatile scenario, namely changes in the network infrastructure and demand, new detection stations or failure ones, among others. The framework is also able to infer automatically the most relevant features to our end-task, including the relationships within the road network. Although the intention with the proposed framework is to evolve and grow with new incoming big data, however there is no limitation in starting to use it without any prior knowledge as it can starts learning the structure and parameters automatically from data. We test this predictive system in different real-work scenarios, and evaluate its performance integrating a multi-task learning paradigm for the sake of the traffic prediction task.

1. Introduction

Nowadays the trend for short-term traffic forecasting relies on data-driven empirical approaches, given the growing data availability, referred to as a big data. This creates the necessity to handle both structured and non-structured data as well as to take advantage from contextual information and data coming from multiple sources and observation technologies. The term short-term traffic forecast in this paper refers to ability of the methods to predict traffic in 15 up to 60 minutes ahead by handling historical traffic data with real-time traffic streams continuously. This means that current and future traffic conditions must be estimated and/or predicted at any point in time, at least 15 minutes ahead, in as short time as possible, based on the most up-to-date traffic data. Additionally, the short-term traffic forecasting task is inherently a real-time task that must deal itself with the usual challenges found in this field, namely high-dimensionality and non-linearity, noisy data from the measurement devices, missing data from faulty or disabled ones, volatility, and adaptation to change in the traffic demand and the traffic supply characteristics. For these reasons, it is widely accepted that a non-parametric approach is usually required to manage the growing complexities as new data is collected.

To deal with some of these difficulties, shallow neural networks and, more recently, deeper architectures have been applied extensively in the short-term traffic forecasting field as they are considered well suited to problems where (i) the input–output data are noisy; (ii) the relationships between these variables are multivariate and highly nonlinear; and (iii) the mapping or relationship is poorly understood (van Lint and van Hinsbergen, 2012). In addition, they are well suited for online learning with new incoming data as there are very well studied optimization techniques such as stochastic gradient descent (SGD) which can be applied to tune the parameters over time (learning), in fact the online optimization field is also an active research area nowadays. Besides the usual huge time required to train deeper architectures, the main drawback of this approach is the lack of interpretability and causality in the results, because often the traffic manager agent does not only care about the final accuracy results, but also about understanding the factors that mostly influenced such results. This is usually not possible with neural networks as they work as a black-box approach. In addition, other works reviewed in the literature simply disregard this kind of problematics and set up experiments with cleaned, imputed and even sometimes dropping out anomalous samples (e.g. holidays) from the testing datasets; these scenarios are far away from the expect in a real-time operating setting.

The paper is organized as follows. In the first part of the paper we will outline the developed autonomous approach for short-term traffic prediction and explain techniques and their main properties used in each of the framework components. In the second part of the paper we will demonstrate traffic forecast framework in two real network examples. First experiment demonstrates the performance of the proposed autonomous approach versus a more classical approach where the forecasting models are built with historical data and updated with new data after an established periodical timescan. The second experiment is designed to demonstrate the advantage of integrated multi-level learning to speed up the traffic prediction by reducing the number of rules. The paper closes with a discussion on further application perspectives of developed autonomous framework and implications for professional practice.

2. Methodology

For the sake of a robust short-term traffic forecasting methodology, we develop a framework built from different machine learning and data analysis components whose predictive system is robust to outliers, irrelevant features and missing data. Developed framework is scalable in terms of network size and can handle growing modelling complexity with new data arrival and adapt to changes in traffic conditions through concept drift detection. The framework is inspired by the works of (Gama, 2010) applied to data streaming scenarios, but tailored to the requirements for this application. In the following sections, the different components of developed framework are presented.

The framework for traffic prediction, Adarules (Mena-Yedra et al., 2017), works in a supervised manner, meaning that for each desired prediction target, i.e. different network locations or forecasting horizons, it is going to discover or unveil a set of rules to gain knowledge about the supervised task, having past observations with their correct prediction. Then, each rule R contained within each ruleset $\mathfrak{R}$ is composed of an antecedent A and a consequent C with the logical form: $A \Longrightarrow C$. The rule antecedent can be composed of several literals L, where a literal L is a single condition over a specific attribute with a specific split-point v; with the form $(xj>v)$,$(xj\leq v)$ if it is numerical, or $(xj=v)$ if it is categorical. $L(xi)$ returns True if xi satisfies L, and False otherwise. The antecedent is interpreted as a conjunction. In this way, a rule R is said to trigger, or to cover, an example xi if all its literals (the antecedent) are evaluated to True on the example

The consequent of a rule is composed of an adaptive output using the multiple rule predictors that the rule may hold (e.g. constant, weighted mean, linear model, or any other functional form). The individual outputs are built at prediction time from the examples gathered in the scope of that rule, then the adaptive output is generated from that population of individual outputs (also could be called experts, following an expert advice schema) weighted by their respective online errors. In addition to the prediction point estimate, an uncertainty interval is given based on the error seen which approximates the real one as the uncertainty associated with covariates is neglected. Finally, each rule R has an associated data structure $\mathcal{L}$ which contains updated statistics from the observed streams (attributes, targets and errors) for those observations gathered by the rule. These statistics are later used for multiple aspects: making predictions, detecting distributional changes and anomalies, evaluating the expansion of a rule, etc. The framework has been designed and implemented based on a modular architecture as presented in Figure 1 such that each unit can be separately replaced or improved. Each component of this framework is described in this section.

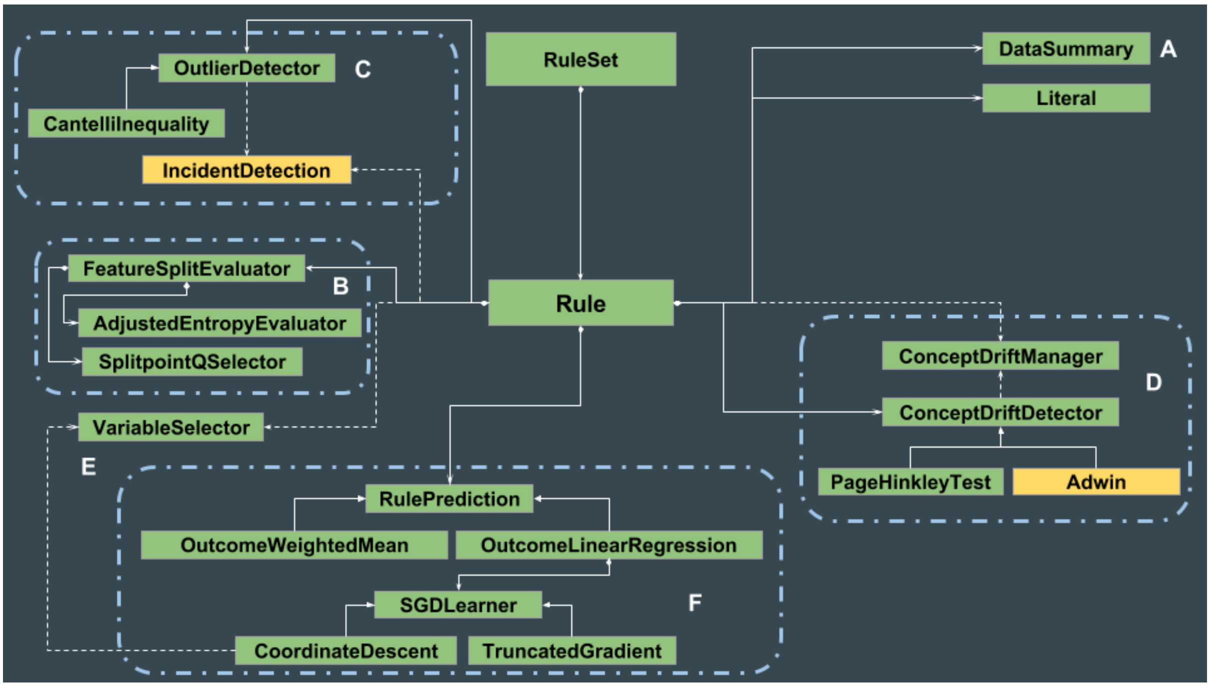

Figure 1: Framework units' graphical schema. Units in green are implemented and in production. Units in yellow are considered for future integration.

The traffic prediction framework consists of the following components: the variable selector (E), the anomaly (outlier) detection (C), the change detection (D), and the handling of missing data and winsorizing which take place inside the data summary structure (A).

Component E: Variable selector to manage prior knowledge about the road-network

The variable selector is the unit aimed to handle the prior information usually put as expert knowledge. In our case, it just set a normalized attractiveness value to each feature separated in different categories (e.g. count, occupancy, speed, time, weather…). Thus, they have an associated probability, that could be updated with new gathered evidence, that is used to select features stochastically as will be explained later in the rule expansion process and in the online learning procedure. Therefore, features associated with detectors in the road network have a normalized attractiveness based on their distance to the point to be predicted. More specifically, the attractiveness is set by the function 1d⁄, where d is the orthodromic distance which could be easily replaced by using travel times coming from a transport network model. On the other hand, discrete attributes (e.g. time, weekday, weather…) have a uniform probability in the scope of its own category to reduce the computation time stochastically. Anyway, all this kind of prior knowledge can be adjusted manually beforehand, or a function can be set to adjust these probabilities in runtime.

Component C: Anomaly detection

Detection of outliers or anomalous examples is very important in on-line learning because of its potential negative impact in the performance of the learning process. For this reason, incoming samples are analysed to detect anomalous samples and to avoid its learning. The current approach is based on the Cantelli's inequality (Bhattacharyya, 1987).

Component D: Change detection

Change detection, also known as concept drift detection in the machine learning community (Gama et al., 2014), is a critical component for modelling non-stationary processes as it is our case. For this purpose, each rule has associated a change detector which monitors their error. The idea is that, after a rule has been expanded and, thus, two new rules are created: their individual rule predictors are trained in their respective ‘batch’ mode with their corresponding gathered observations. From this moment, their residual mean error should be located at zero and it is started to be monitored for changes. When a change is detected (i.e. a significant increase in the error), a signal is sent to the concept drift handler and the rule is removed from the ruleset. The current implemented approach for detecting a change is based on the Page-Hinkley (PH) test (Page, 1954), although other approaches are being considered (Bifet and Gavaldà, 2009, 2007).

Component A1: Handling missing data

The framework gathers online statistics for each attribute in the context of each rule which corresponds to specific road conditions. So, in the long term, with enough sample size each rule has a good view of their data distribution for each recognized road condition. Thus, for each missing attribute, the framework reconstructs a normal distribution with the gathered mean and dispersion, but limiting the probability density at zero at the current minimum and maximum values in order to avoid extrapolation in the covariates. Finally, missing values can be replaced with samples gathered from this distribution.

Component A2: Winsorizing for extreme values

Winsorizing is a statistical technique to filter extreme values. When extreme values (outliers) in traffic data from sensors are received, i.e. those whose probability is extremely low in the scope of a specific rule, it is often better to filter them or else replace them using the handler for missing data described above. In result, if we assume traffic variables are modelled as Gaussian distributions. In this method, we have assumed Gaussian distribution of traffic variables, such that those values beyond or above approximately 3 standard deviations from the mean value are treated as a spurious outlier.

2.1 Rule expansion

Rules could be viewed as high-level features, i.e. patterns of traffic parameters, discovered in the road network with the aim of reducing the uncertainty around the prediction target using a specific goodness of fit function. For this purpose, existing rules have a chance to run a rule expansion evaluation process (component B in Figure 1). If the evaluation process is favourable, the current rule disappears and it is specialized into two new rules with their respective observations and statistics. The frequency of this evaluation, which takes place for each rule separately, is crucial as a low frequency can lead to a slow learning of the high-level features while a high frequency can make the process too sensitive to transient noise. The parameter $Nmin$ dictates the minimum amount of observations which must be seen, separately on each rule scope, to proceed with a rule expansion evaluation. This threshold $Nmin$ is pre-set to an initial value $Nmin0$, that is later dynamically adjusted, but never increasing, based on the dispersion of the rule error in a logarithmic scale. This dynamic adjustment aims at relaxing the trade-off between prompt but expensive checks and slow but inefficient checks. A high initial value can be set because, afterwards, it is going to be adjusted automatically based on the dispersion of the error rule, which means that if the rule is having a narrow error then it is not necessary to try to specialize it so often. In the end, if the rule expansion evaluation process is successful, expanding a rule R consists of creating two new separate rules ($Rleft$, $Rright$) with their respective observations by adding the new literal created with the corresponding attribute and split-point to the sets of antecedents.

There are two steps in the rule expansion evaluation process, namely: (1) the searching step to find which attributes along with their corresponding split points are going to be evaluated, and (2) the scoring process to rank those selected combinations.

1. Reducing the search for rule expansion

When it is time to run the rule expansion evaluation process, it is needed to decide which attributes and split points are going to be measured. Perhaps the intuitive idea is simply to evaluate all the existing features, but in the current high-dimensional problem this can lead to time-consumption problems especially if the threshold $Nmin$ is low. Not only that, overfitting may occur if, for instance, detectors that are very far away are selected as antecedents. Therefore, the candidates to be evaluated are selected probabilistically based on their distance using the variable selector.

The split points to be evaluated for each selected continuous attribute, are selected using the cumulative probabilities, or quantile functions, to represent the whole distribution of the gathered observations. While in the case of discrete attributes the selection is based on the generation of multiple continuous intervals.

Continuous attributes considered include the traffic count, occupancy and speed from the whole road network. Discrete attributes considered include the time of the day, weekday and weather information.

2. Scoring the candidates for rules' literals

The goodness of fit used to evaluate the different combinations of features and split-points is based on entropy minimization, process which is also known as information gain. From an information theory perspective, entropy $H(X)$ measure the randomness of the information in the random variable X. The entropy is maximized if the distribution is vague (i.e. uniform with equal probability in the whole space), this is the situation of maximum uncertainty as it is most difficult to predict the outcome. When there is less uncertainty, i.e. when the outcome is peaked around certain location values, there is a lower entropy quantity. At the extreme case, when there is no uncertainty because we are sure about the outcome the entropy is zero (MacKay, 2003).

When scoring a proposed splitting, entropy is used as information gain score. This means that we score the entropy of the current rule before splitting versus the entropy of the proposed new rules weighted by their respective new sample sizes. If entropy is reduced with the new splits, that means we have gained certainty about the outcome.

In addition, the goodness of fit function considers the missing data ratios of the feature candidates, penalizing those whose missing data ratio is higher considering these as untrustworthy candidates.

2.2. Rule prediction

Currently, there are two proposed strategies to forecast within the rules' scope, and a strategy to combine these forecasts into a single point-estimate prediction.

1. Weighted mean

This forecaster is simply the weighted historical mean of the true target of the past examples covered by the rule. This is equivalent to a naïve predictor, which is good to maintain among the forecasters population as it has no direct dependencies on external states.

2. Penalized linear regression

A linear regression model is built using the examples covered by the rule. Although short-term traffic prediction is a highly non-linear problem, we use the rules to discover the nonlinearities and combine a population of lower-level, specialized linear models.

3. Adaptive strategy

Finally, an adaptive strategy combines the forecasters population derived from the previous two strategies that exist within a rule, namely: the weighted mean and the different penalized linear regressions. This adaptive strategy is based on the on-line estimation of the mean absolute error (MAE), where the contribution from each forecaster to the final point-estimate prediction is determined inversely proportional to their current online estimation of the error.

3. Experiment design and results

3.1 Experiment 1: Forecasting comparison against a classical blind adaptation approach

In this experiment, we want to assess the performance of the proposed autonomous approach versus a more classical approach where the forecasting models are built with historical data and updated with new data after an established periodical time.

The data used in this paper comes from the Caltrans Performance Measurement System (PeMS) maintained by the California Department of Transportation (California Department of Transportation, n.d.). More specifically, the current work has focused the attention into the Caltrans District 11 with over 1,500 detection stations corresponding to the City of San Diego (US), Figure 2. Collected data for experiments spans for three years ranging from 2013/01 to 2015/12 with an initial 5-min resolution with has been aggregated to 15-min for three reasons: (a) mitigating the inherent noise in road network measuring devices, (b) reducing the running time for the experiments in the current research work without compromising the validity of the results, and (c) convenience for commercial purposes from TSS-Transport Simulation Systems and its product Aimsun Online. As the final intention is that the resulting output from this research can be used in an online prediction system, we have focused on predicting traffic volume because of its use in the matrix estimation process which consists in estimating, from individual link flows (and turning proportions), an aggregated demand matrix (OD flows) which serves as input for the simulation step.

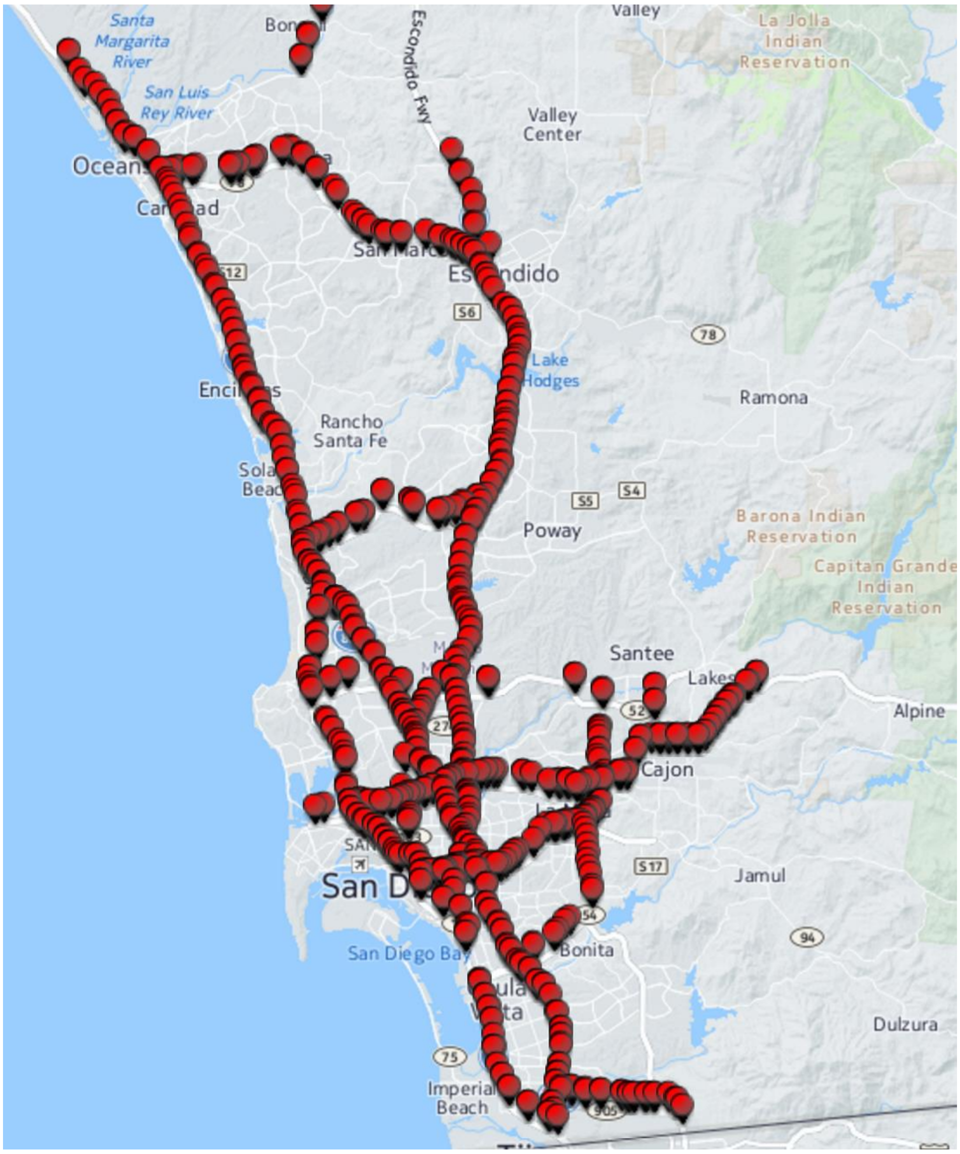

Figure 2: Induction loop detectors (+1500) in the network of San Diego.

For the experiments, a subset of detection stations from the road network has been chosen to focus on a small set of locations which are far apart from each other, but exhibit high temporal variability. These can be observed in Figure 3, where an exploratory analysis shows the distributional changes over time for the six selected detection stations. Data (y-axis) corresponds to traffic volume [vehicles / 15 minutes] with distributions (x-axis) corresponding to monthly aggregates (2013/01 to 2015/12). The coloured boxes represent the interquartile range (25 to 75% of data within), black thicker line stands for the mean and black thinner lines stands for minimum and maximum during the period.

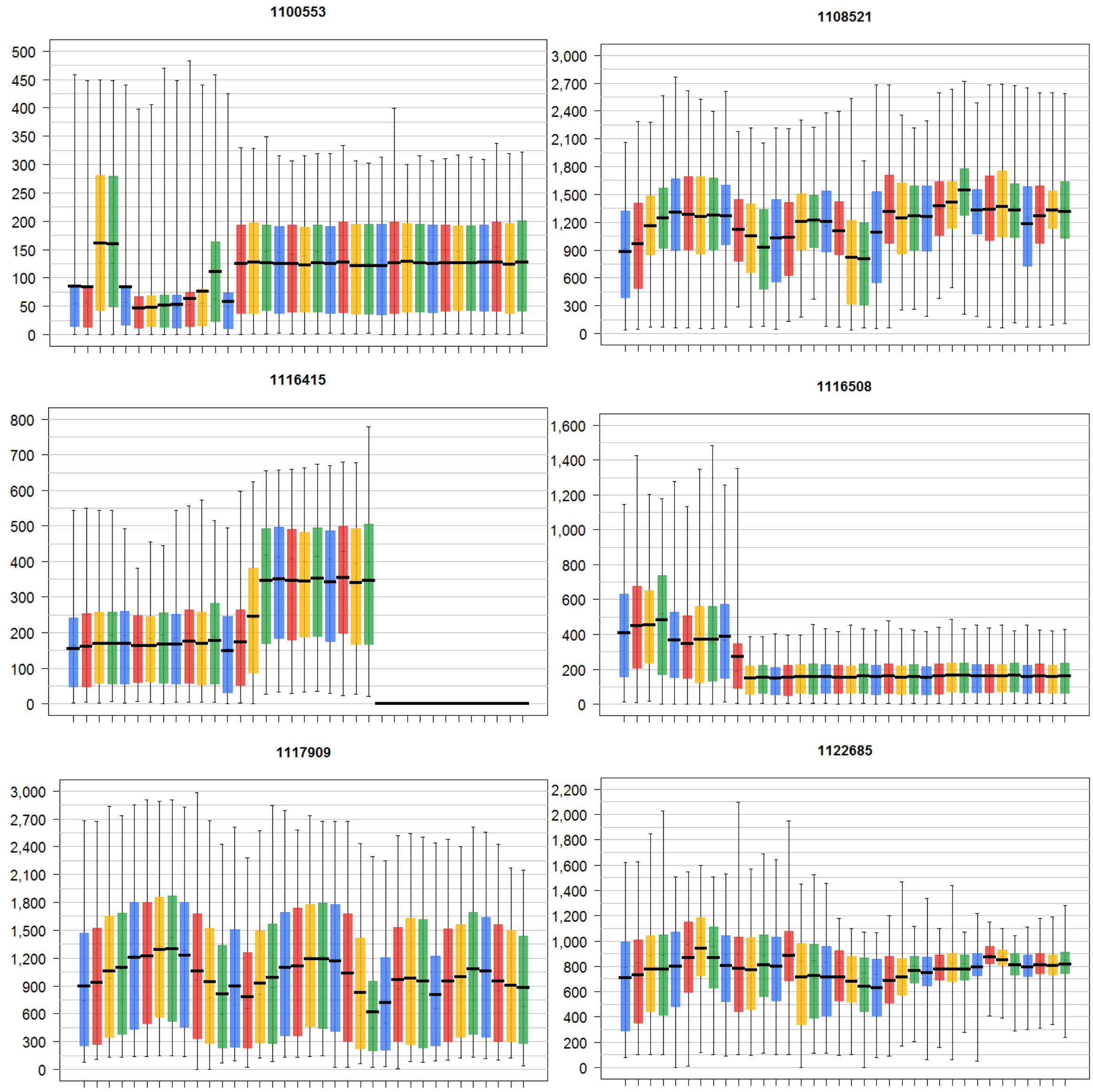

Figure 3: Exploratory analysis showing the distributional changes over time for the six selected detection stations. The y-axis represents the monthly traffic volume [vehicles / 15min] distribution. The x-axis represents the time, with each bin corresponding to a month.

In addition, the goal of our experiments is to compare the performance of our real-time learning framework with the alternative modelling approaches:

1. classical approach of training a model in batch mode.
2. classical approach of training a model in batch mode with blind model's parameter adaptation schedule.

Classic approach of batch training is using only real traffic data to predict traffic flow but not to update the model parameters. The batch approach is based on the same linear regression penalized with L1 (LASSO) as adopted in the developed real-time learning framework. The best λ penalty value is selected using cross-validation on the training dataset, and the input information is composed from the traffic volumes existing in the whole road network at the same time. Although the short-term traffic prediction task is non-linear problem, we have defined it as a liner model, such that non-linear behaviour is modelled by discretising the task according to the data resolution (15 minutes) and training a separate set of coefficients for each time point. Second approach is based on the blind adaptation model's parameter adaptation schedule. Blind adaptation approach is retraining the whole models set every 1 week or 1 month using the last 1 month or 6 months for the training dataset.

Finally, a 60-min forecasting horizon has been chosen to evaluate the different approaches because it is a challenging and interesting horizon for commercial purposes. The performance metric considered to compare results is the mean absolute percentage error (MAPE) because it gives an intuitive measure of the performance independent of the unit and scale, and a time interval of one month has been used to aggregate the MAPE values over.

The results of the experiments are shown in Figure 4. As can be seen, the MAPE for the adaptive approach is initially high in all the stations because the framework starts with no knowledge and then it starts to learn and adapt its parameters as in real-time. On other hand, the approaches based on batch training start with a low error because it is just the data which has been used for their training (the first 6 or 12 months from the 3 years). Obviously, it is not a fair comparison, but the aim is just to show how Adarules lower its error as more data is seen and more knowledge is acquired. It can be seen also that the batch approach trained with more data (1 year) has a lower error than the batch approach trained with less data (6 months) but the difference is slight. Another interesting point is observing how the adaptive approach deals better when sudden changes happen, while the performance for the batch approaches deteriorates and does not seem to recover. It can be seen also that the performance using the adaptive approach is improving until the end of the experiments. This can be easily observed in the figures, because it is when the MAPE value increases. Finally, when it is compared to the blind adaptation approaches; the accuracy performance is similar on the long term, however a crucial difference is the Adarules autonomy to decide the training times avoiding unnecessary training costs every 1 week or 1 month. Besides collateral benefits from Adarules explained in previous section, another crucial difference not noticeable in the picture because of the results aggregation, is that Adarules is giving responses even with missing data.

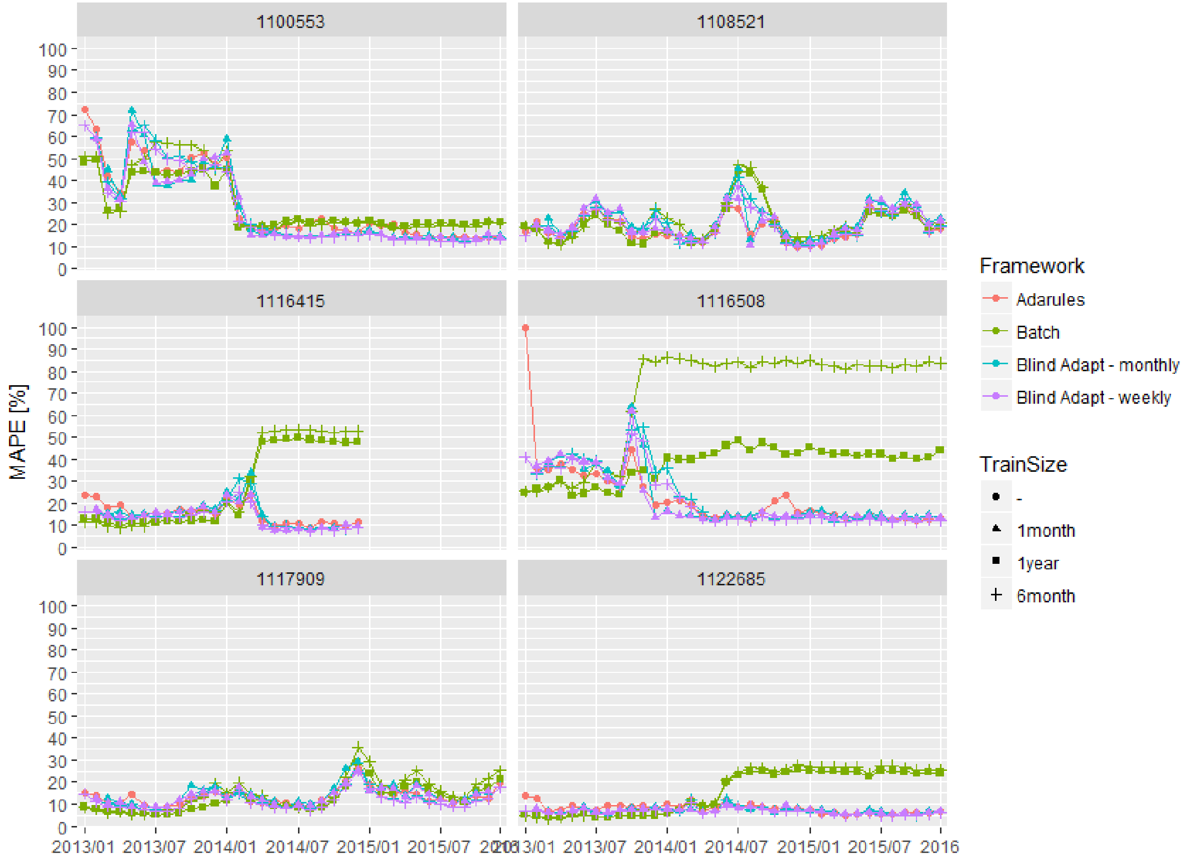

Figure 4: Experiments results for six detection stations and 60-minute forecasting horizon, showing the monthly aggregated MAPE for the three approaches and its evolution over time during the three years.

3.2. Experiment 2: Checking the ability of the system to identify the best spatiotemporal correlations in the network over time

In this case, data from the City of Santander (Spain) is used, an urban network with more than 300 individual inductive loop detectors collecting data in real-time on traffic vehicle counts, occupancies and speeds (). The network has been chosen as the sensors span the entire City, there is a rich information, and given that it is an urban network it makes a challenging scenario with more than 4000 links. Collected data for experiments spans an entire year ranging from January to December 2016 with a 15-min aggregation resolution.

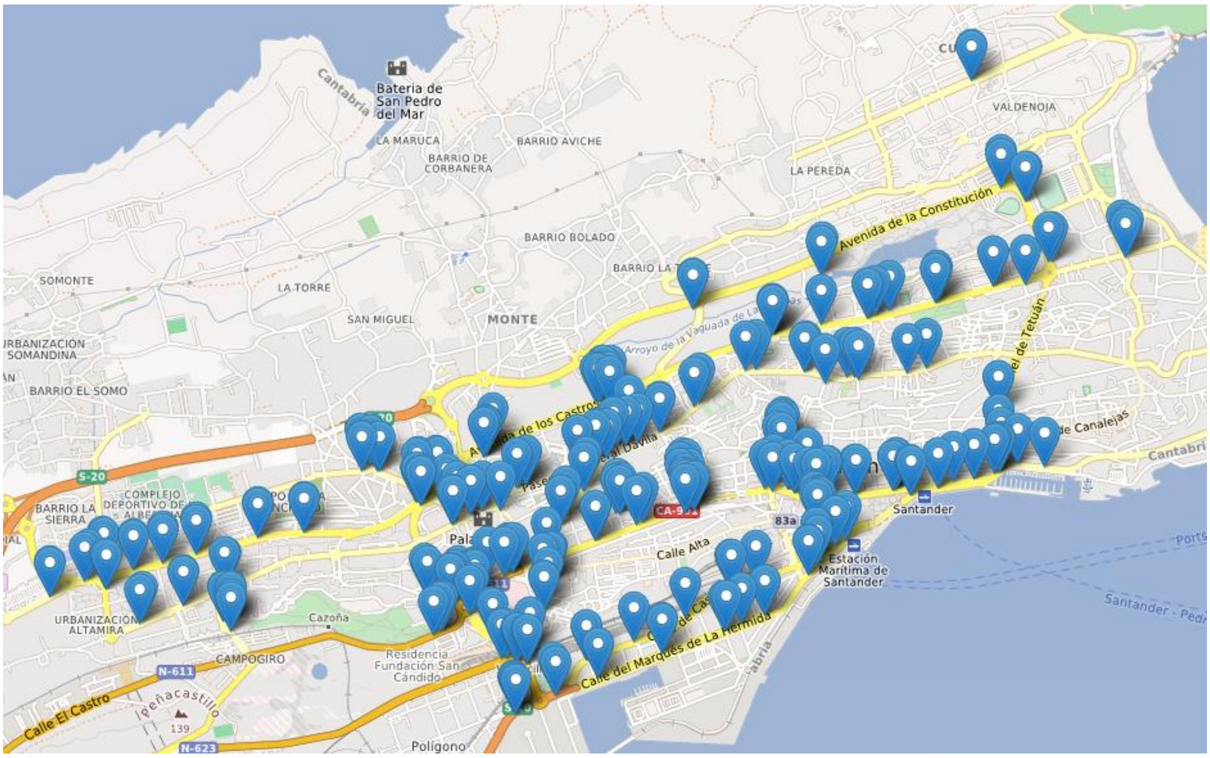

Figure 5: Induction loop detectors (+300) in the network of Santander.

Traffic prediction at local detector is performed by using data collected from local detector and other sensors in the network. The amount of data used from different sensors in the network can have a strong impact on prediction results due to their correlations. For this reason, the performance of the local traffic prediction method is evaluated for various settings of spatial distribution of sensors from the local one. The following eight experiments are defined to evaluate impact of spatial distribution of sensors from the local one (in km) on local traffic predictions: 0.25, 0.5, 1, 2, 3, 4, 5, and the whole network scope. This experiment would test mainly the components B and E in the Figure 1.

Figure 6 shows sensitivity analysis results of the local traffic prediction method on spatial distribution of sensors from the local one, where bars within each monthly slot represent experiments in consecutive order. Sensitivity results are presented as the error interquartile-range [25% to 75%] and the mean for monthly error aggregates depicted by black line, as shown in Figure 6. Results show that relative prediction error decreases with the spatial increase of detectors involved in local traffic prediction. Although prediction relative error decreases over time and reaches a steady state, keeping narrow detection of traffic dynamics to local detector will have a higher impact on the prediction error.

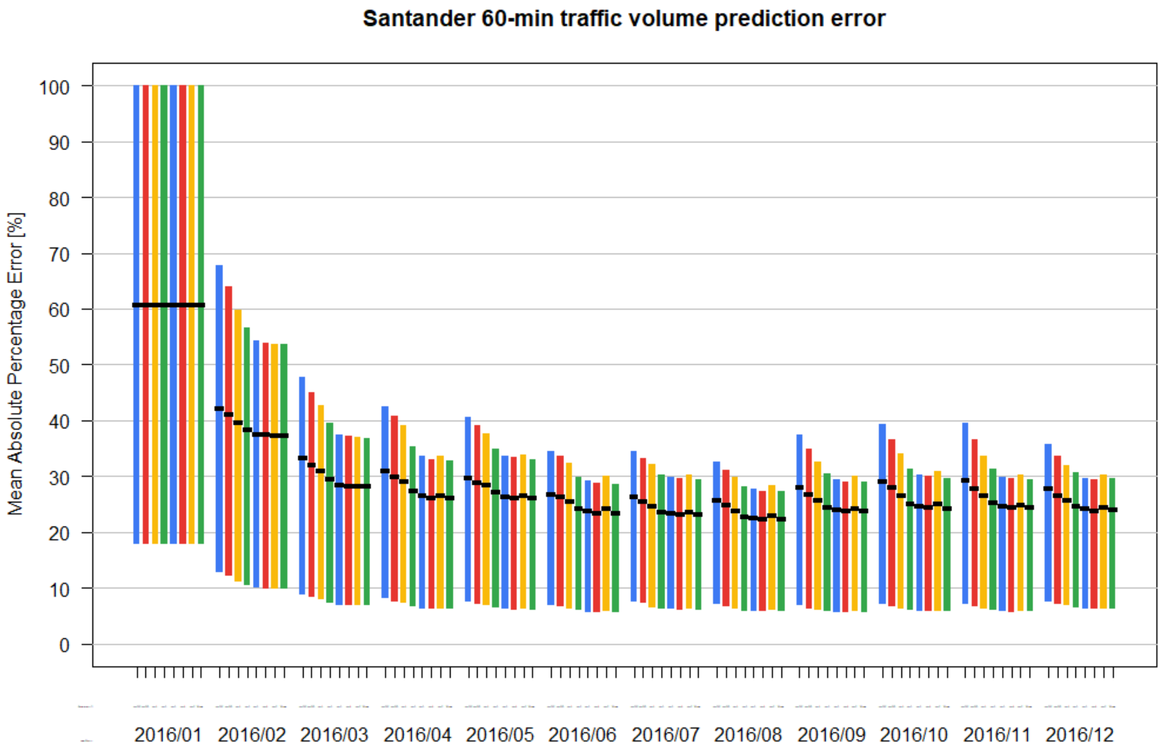

Figure 6: 60-min traffic volume prediction error for the main corridor in Santander.

The performance of the traffic prediction over the one year period in 2016, and for each defined experiment is quantified and summarized for key statistical performance indicators in Table 1. Results show that the best performance of the developed local traffic prediction method will be achieved when method is set to automatically decide which spatial points are more relevant for the predictive task instead of manually limiting the spatial visibility.

Experiment	Min (10^{-14})	1st Qu.	Median	Mean	3rd Qu.	Max.
Km0.25	2.22045	7.73991	17.85171	30.00118	40.794	100
Km0.50	2.22045	7.297	16.8683	28.8923	38.66504	100
Km1	2.22045	6.88416	15.92317	27.86802	36.79545	100
Km2	2.22045	6.51369	15.03571	26.6294	34.42114	100
Km3	2.22045	6.31492	14.59055	26.11157	33.36413	100
Km4	2.22045	6.19712	14.35497	25.82217	32.92857	100
Km5	2.22045	6.32554	14.63455	26.25434	33.6181	100
Whole	2.22045	6.22601	14.37244	25.77954	32.73567	100

Table 1: Traffic prediction reliability for the main corridor in Santander: statistical performance indicators.

Experiment 3: Integrating a multi-task learning approach for traffic prediction

Multi-task learning (MTL) is a paradigm in the realm of machine learning itself. The core idea, as was defined by (Caruana, 1997), is to act as an inductive bias causing a model to prefer the hypotheses best explaining the set of related tasks simultaneously. By doing this, the main goal is to improve generalization performance by leveraging the domain-specific information contained in the training signals of these related tasks. In the end, this approach aims to make the framework more efficient by reducing the number of learnt rules, while regularizing the learning procedure when multiple related tasks are learnt at the same time. This experiment would test mainly the components B and F in the Figure 1.

In our experiments, the used approach tries to leverage the fundamental relations in traffic expressed in the well-known fundamental diagram of traffic flow (Transportation Research Board, 2011) for each specific detector or spatial point in the network. The basic idea is that the proposed predictive system enhanced with the MTL paradigm unveils spatiotemporal correlations in the road network or associated with qualitative variables such as the time to better perform in the current predictive task, while giving an interpretable reasoning of the most influential factors. In the end, the system leverages this jointly learning for each detector to identify traffic conditions, i.e. free-flow, bound flow or congestion.

For the validation of the proposed experiments, data used comes from the City of Santander. For this experiment, a subset of detection stations from the road network has been chosen to focus on a small set of locations. We chose eight detection stations which are far apart with the purpose of achieving a representative picture over the entire road network instead of focusing on a small area; and having two of them (1021, 1023) in one of the main entrance/exit to the City, two of them (3078, 3079) in another main entrance/City, two of them (2016, 2019) in one of the main arterials, and finally another two (2057, 2070) in the City centre, with the purpose of capturing sufficiently different dynamics from the network.

The goal of this experiment is to check if jointly learning both traffic variables that have a well understood relation defined in the fundamental diagram of traffic flow helps to the predictive task by letting the system identifying these situations (free flow, bound flow, congestion) through rules discovery in the data. To this aim, we have learnt different rulesets for each detector. On one hand, we have learnt a ruleset to predict the traffic flows on each location and another ruleset to predict the respective occupancies, which means 16 rulesets overall. On the other hand, we have learnt a ruleset for each detector for both variables at the same time, which means 8 rulesets.

Results for 60-min traffic flow prediction can be seen in Figure 7. In this figure, a more aggregated view of all the detectors performance is shown, but it is easier to check if there is a significative difference in the cumulative percentage of the MAPE below a certain value. It can be observed that it is evident that the prediction performance is practically the same during the four periods. The Empirical cumulative distribution function (ECDF) plots show the same curve which means there is no significant difference in the traffic flow prediction performance between learning solely traffic flows or jointly learning both variables together. This is confirmed in Table 2, where more disaggregated results for each detector corroborate that almost identical MAPE values are obtained for both approaches. More specifically, the averaged MAPE for both approaches are around the 15% for all the detectors and periods excluding the off-peak period that obviously has a higher averaged MAPE, around 40%, due to the low traffic volume and its impact on this relative metric. However, it is interesting to note that one of the detector with highest error using the single-task approach (2019) obtains a slight reduction of almost a 2% using the MTL approach for the morning, noon and evening periods.

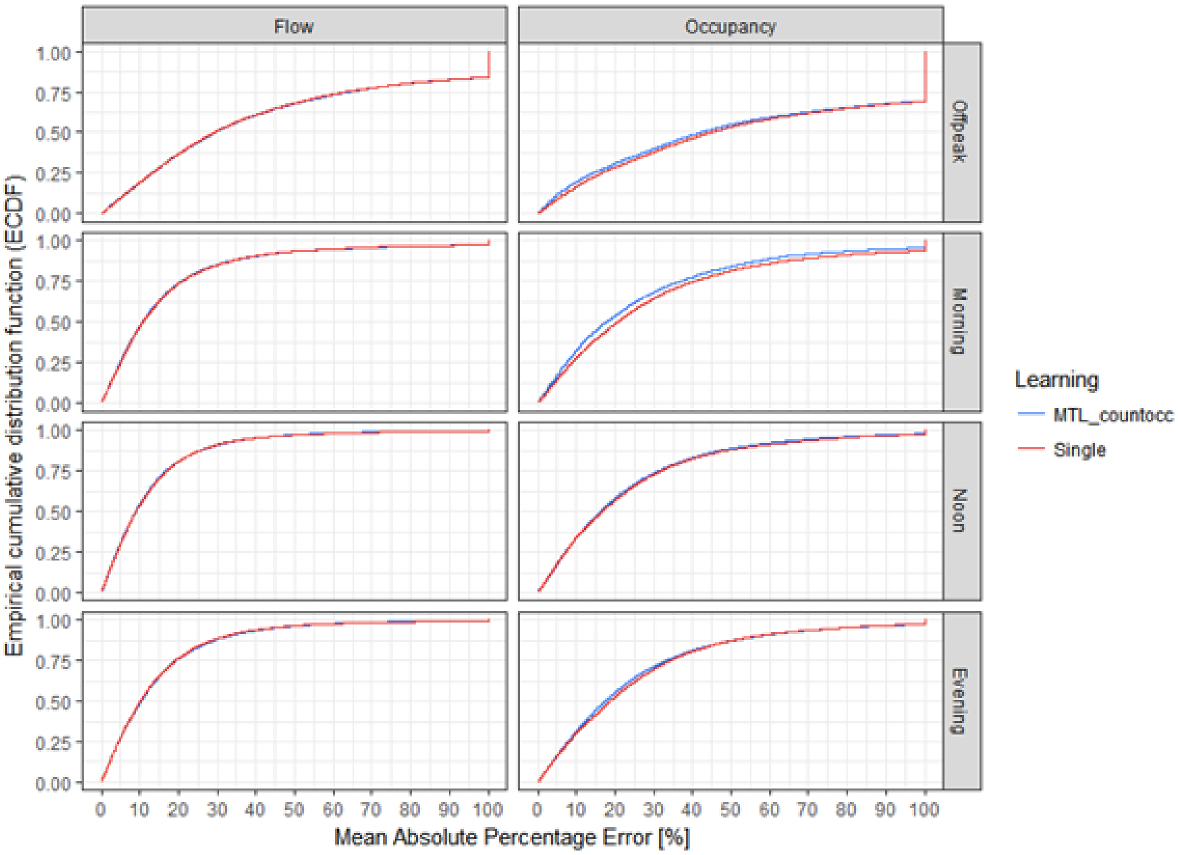

Figure 7: ECDF showing the cumulative distribution of the MAPE predicting traffic flow and occupancy for each period.

Predicting occupancies is clearly a harder task due to the small scale, the high non-linearity and the presence of noisy data or even binned data. For that reason, higher MAPE values are obtained compared to predicting traffic flows. In this case, there is a tiny improvement in all the periods, but especially in the morning period with a decrement of around 3% in the averaged MAPE.

Flow	Off-peak		Morning		Noon		Evening		No. of rules	
Source	**Single**	**MTL**	**Single**	**MTL**	**Single**	**MTL**	**Single**	**MTL**	**Single**	**MTL**
1021	22.36	22.37	10.73	11.84	6.99	6.94	7.43	7.39	22	23
1023	41.25	41.50	20.81	21.83	11.04	9.97	11.06	12.47	14	23
2016	32.65	32.83	12.62	13.03	10.07	10.72	11.58	12.04	24	23
2019	56.61	57.22	25.20	24.04	23.29	21.72	26.15	24.28	28	27
2057	39.23	39.58	18.60	19.11	14.76	15.68	17.20	17.91	25	23
2070	37.98	36.67	16.42	16.10	11.88	11.83	12.59	13.04	21	23
3078	44.65	45.70	15.75	15.32	12.58	12.61	14.30	14.57	18	21
3079	52.43	52.74	17.75	16.12	13.88	13.66	17.05	16.58	17	19
Mean	**40.67**	**40.79**	**17.21**	**17.13**	**13.08**	**12.91**	**14.67**	**14.79**	—	
Occupancy	**Off-peak**		**Morning**		**Noon**		**Evening**		**No. of rules**	
Source	**Single**	**MTL**	**Single**	**MTL**	**Single**	**MTL**	**Single**	**MTL**	**Single**	**MTL**
1021	24.86	20.44	15.49	15.31	11.93	11.31	13.21	11.47	9	—

1023	42.45	38.34	26.04	24.05	15.38	15.87	20.80	19.12	3	
2016	49.53	46.48	41.72	36.44	38.10	35.31	41.18	38.00	9	
2019	56.08	57.04	26.20	23.89	21.43	21.74	24.72	21.94	8	
2057	63.27	64.75	40.59	39.64	32.77	32.80	29.98	30.73	11	
2070	55.11	55.21	25.73	25.75	21.70	21.94	25.88	27.87	14	
3078	68.37	67.80	19.94	17.67	19.05	18.23	21.27	19.64	6	
3079	70.11	68.94	38.56	27.41	28.77	26.83	24.43	27.89	7	
Mean	**52.83**	**51.43**	**29.35**	**26.33**	**23.65**	**23.01**	**25.18**	**24.58**	—	
Total	—								**236**	**163**
Median									32	23

Table 2: Statistics showing the averaged MAPE predicting traffic flow and occupancy for each period. Results are shown for each detector and aggregate.

Besides the comparison of the prediction performance, which is a critical aspect of any predictive system, there is another key factor regarding the interpretability and efficiency of the system. Interpretability was one of the aim this traffic prediction system was built for, thus any factor that can improve this aspect is especially important for traffic engineers and managers that use this tool. As can be observed in Table 2, the number of rules is reduced noticeably in the case of using the MTL paradigm, going from a total of 236 to 163 rules, and going from an average of 32 rules per ruleset to an average of 23. This is done without sacrificing performance, or even improving it at some points.

A final remark is that Santander is not a City with a high presence of congested links so probably this kind of study should deserve another analysis in a kind of network with a higher presence of congestion events.

4. Conclusions and Implications for Professional Practice

The experimental results have confirmed the expectations about the proposed Adarules, which are a good tolerance and fast adaption to change, especially sudden changes and long-term changes associated with seasonality or traffic demand growth. Second, that as it sees more data, the framework learns and maintains its predictive accuracy. Third, it is valuable that the framework unveils the inherent dependencies in the road network (which can be seen as high-level features in the machine learning argot) and, also importantly, these can be easily interpreted and evaluated by traffic managers. The use of contextual information (e.g. date, time and weather) and the measurement of its impact is especially attractive for traffic managers. Another interesting advantage of Adarules is that it can give responses for the predictive point even if it is temporally malfunctioning if there is enough knowledge acquired and thus reconstructing its typical behaviour. Finally, Adarules is an autonomous framework and can manage the trade-off for deciding the proper training times to adapt quickly to changes while keeping a good prediction accuracy. Another kind of experiment tested the integration of a multi-task learning approach for the sake of forecasting both traffic flow and occupancy. More specifically, the first experiment tested a single-task learning for traffic flow and occupancy prediction separately and a multi-task learning approach that jointly learns both. The results showed that there was no significant improvement in the traffic flow prediction performance, and only a slight improvement in the occupancy prediction task. Efficiency and interpretability were improved by reducing 40% of the rules created in the single-task learning approach. However, this requires more testing in other networks with more congestion events.

Finally, an important remark about road network traffic prediction is that prediction accuracy is very important, but it cannot be the only criterion when choosing the appropriate modelling methodology (Kirby et al., 1997). Given that this task is a non-stationary stochastic process tackled in real-time, other matters concerning the adaptability to changing behaviours and traffic demand changes, transferability to new locations with scarce data or information about the traffic supply characteristics, causality and interpretability about the process, and cost in time and effort for model development and, more importantly, maintenance must be considered. Some of these challenges are pointed out in (Vlahogianni et al., 2014), which can be summarized in the following points that we believe that the proposed predictive framework Adarules in the current research work aims to deal with them: (1) responsive forecasting schemes for non-recurrent conditions, (2) developing prediction systems with increased algorithmic complexity, (3) attempting to understand data coming from novel technologies and fuse multi-source traffic data to improve predictions, (4) the applicability of artificial intelligence (AI) methodologies to the short-term traffic prediction problem.

References

Bhattacharyya, B., 1987. One sided Chebyshev inequality when the first four moments are known. Communications in Statistics-Theory and Methods 16, 2789–2791.

Bifet, A., Gavaldà, R., 2009. Adaptive Learning from Evolving Data Streams, in: Proceedings of the 8th International Symposium on Intelligent Data Analysis: Advances in Intelligent Data Analysis VIII, IDA ’09. Springer-Verlag, Berlin, Heidelberg, pp. 249–260. doi:10.1007/978-3-642-03915-7_22

Bifet, A., Gavaldà, R., 2007. Learning from Time-Changing Data with Adaptive Windowing, in: Proceedings of the 2007 SIAM International Conference on Data Mining, Proceedings. Society for Industrial and Applied Mathematics, pp. 443–448.

California Department of Transportation, n.d. Caltrans PeMS [WWW Document]. URL http://pems.dot.ca.gov/ (accessed 1.11.17).

Caruana, R., 1997. Multitask Learning. Machine Learning 28, 41–75. doi:10.1023/A:1007379606734

Gama, J., 2010. Knowledge discovery from data streams. CRC Press.

Gama, J., Zliobaite, I., Bifet, A., Pechenizkiy, M., Bouchachia, A., 2014. A Survey on Concept Drift Adaptation. ACM Comput. Surv. 46, 44:1–44:37. doi:10.1145/2523813

Kirby, H.R., Watson, S.M., Dougherty, M.S., 1997. Should we use neural networks or statistical models for short-term motorway traffic forecasting? International Journal of Forecasting 13, 43–50.

MacKay, D.J., 2003. Information theory, inference and learning algorithms. Cambridge university press.

Mena-Yedra, R., Gavaldà, R., Casas, J., 2017. Adarules: Learning rules for real-time road-traffic prediction. Presented at the 20th EURO Working Group on Transportation Meeting, EWGT 2017, Budapest, Hungary.

Page, E., 1954. Continuous inspection schemes. Biometrika 41, 100–115.

Transportation Research Board, 2011. 75 Years of the Fundamental Diagram for Traffic Flow Theory: Greenshields Symposium, Transportation research circular. Transportation Research Board, Woods Hole, Massachusetts.

van Lint, H., van Hinsbergen, C., 2012. Short-Term Traffic and Travel Time Prediction Models. Transportation Research E-Circular.

Vlahogianni, E.I., Karlaftis, M.G., Golias, J.C., 2014. Short-term traffic forecasting: Where we are and where we’re going. Transportation Research Part C: Emerging Technologies, Special Issue on Short-term Traffic Flow Forecasting 43, Part 1, 3–19. doi:10.1016/j.trc.2014.01.005