Please cite this paper as follows:
Li, Z., Zheng, Z., Washington, S., 2019. Short-term traffic flow forecasting: a component-wise gradient boosting approach with hierarchical reconciliation. IEEE Transactions on Intelligent Transportation Systems. doi:10.1109/TITS.2019.2948381

Short-term traffic flow forecasting: a component-wise gradient boosting approach with hierarchical reconciliation

Zili Li, Zuduo Zheng, and Simon Washington

***Abstract*—A gradient boosting procedure in combination with hierarchical reconciliation is proposed in this study for short-term forecasting of traffic flow. Particular attention is paid to three main characteristics of traffic flow: the temporal and spatial patterns, interactions between the temporal and spatial patterns, and the dynamics of traffic flow at different spatial aggregation levels. The performance of the proposed forecasting framework is examined by comparing it with three frequently used methods (i.e., SARIMA, Kalman filter model and random forest) in the literature, and using three distinctive datasets. Overall, the gradient boosting based approach offers a highly flexible and automated way to learn useful information in large datasets, which is particularly advantageous for forecasting traffic flow in a complex road network at longer forecasting horizons.**

***Index Terms*—Traffic volume forecasting; Gradient boosting; Hierarchical reconciliation.**

I. Introduction

The quality of short-term traffic flow forecasts is crucial to the effective planning and development of modern traffic management and control systems. Accurate short-term forecasts can not only inform travellers about the future traffic conditions, but also help operators to develop proactive traffic management strategies. For this reason, short-term traffic flow forecasting has attracted a lot of attention in the literature [e.g., 1, 2, 3, 4].

Volume, speed and density are three commonly-used measurements for representing traffic states [5, 6, 7, 8]. The approaches for forecasting these quantities can be roughly divided into two main categories: the time series modelling approach, which utilises the strong seasonal pattern in traffic flow and fully explores the autocorrelation in traffic data [9, 10, 11]; and the machine learning approach, which relies on highly flexible semi- or non-parametric frameworks to extract important patterns in traffic data [12, 13, 14, 15].

Despite the different measurements of traffic flow and the different forecasting frameworks used in the literature, there are, however, two fundamental features that are common to all traffic datasets: seasonality [16, 17, 18] and the spatial correlations of traffic flow among adjacent locations [10, 7]. Obviously, successful forecasting of traffic flow requires the model to be sufficiently flexible in accommodating these features and potentially the interactions between the two [19, 20, 21]. The model proposed in this study follows this general principal. Nonetheless, there are two features in the proposed model that are unique compared with others in the literature. First, the forecasts for one location can be adjusted based on the forecasts for other locations and also the forecasts for groups of locations at different spatial aggregation levels; Second, the model is built based on a machine learning technique, gradient boosting [22], and in combination with a novel application of the hierarchical reconciliation of [23]. As the results, multiple seasonal patterns, dynamic spatial correlations among many locations and information on different aggregation levels can be combined and utilised effectively in a single modelling framework for forecasting purposes.

Overall, the proposed forecasting framework is highly flexible and can be applied to scenarios with different degrees of complexity, e.g., it can be used to forecast traffic flow for a single location, for a road segment, or for a complicated urban network. Such flexibility fills a significant research gap in the literature: Most of the existing forecasting methods are only suitable for a single location or a road segment. Alongside the emergence of connected vehicles (i.e., vehicles that can talk to each other as well as to roadside infrastructures), the proposed forecasting framework with the capability of dealing with a complex road network is expected to become increasingly useful because of its strength in utilising rich network information for deriving more effective operational and control strategies.

The remaining of this paper is organized as follows: Section II contains a preliminary data analysis that illustrates the main characteristics of traffic flow. Section III describes the gradient boosting based forecasting framework. In Section IV, the performance of the proposed forecasting approach is examined using a dataset collected in the city centre of Adelaide, Australia. Section V compares the forecasting performance of the proposed approach with that of three existing models using three different datasets. Finally, Section VI summarises the main findings of this study.

II. An empirical investigation on the spatio-temporal characteristics of traffic flow

To explore the main characteristics of traffic flow in a complex road network, volumes (i.e. vehicle counts at a certain time interval) collected at 81 intersections in the city centre of Adelaide, Australia from 2010-01-01 to 2014-06-19 are used. Figure 1 shows the hourly volumes of a week and

(Corresponding author: Zuduo Zheng.)
The authors are with the School of Civil Engineering, The University of Queensland, Brisbane Queensland 4072 Australia (e-mail: zili.li@uq.edu.au, zuduo.zheng@uq.edu.au and s.washington@uq.edu.au).

hourly volumes of a day for one of the intersections (site 3003), averaged over the entire sample period. The diurnal and weekly patterns of the traffic volume can be seen clearly from the figure. In the lower panel of Figure 1, the volume reaches its lowest point in the early morning of a day, increases dramatically at around 8 a.m. before reaching the daily peak at around 9 a.m., and remains at a relatively high level until late afternoon at around 7 p.m. The weekly pattern of traffic volume is also apparent from the upper panel of Figure 1. The volume profile during weekdays is quite similar, and different from that on weekends. These seasonal patterns of the traffic volume are not surprising, and are well documented in the literature [24, 25, 10, 20].

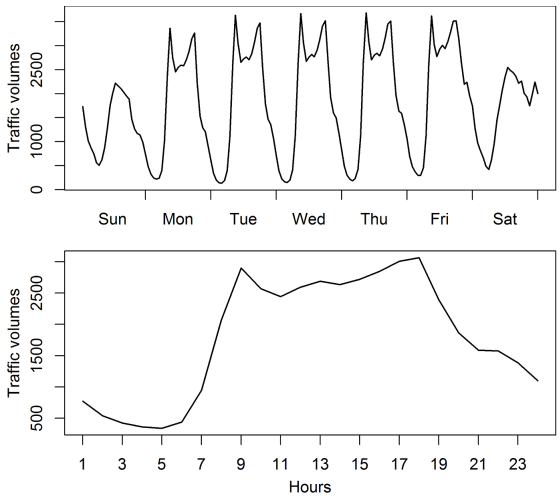


Fig. 1: Averaged hourly traffic volumes of a week and averaged hourly traffic volumes of a day in the upper and lower panels, respectively, at site 3003 in the city centre of Adelaide, Australia.

The seasonal pattern discussed above entails the important dynamics of traffic volumes in the temporal dimension that every forecasting model should consider. In the other dimension of interest, the spatial dimension, it is reasonable to postulate that the traffic volumes at nearby locations are correlated. In other words, when forecasting the traffic volume at one specific location, the volumes at nearby locations may carry important information for inferring the future volume at that location.

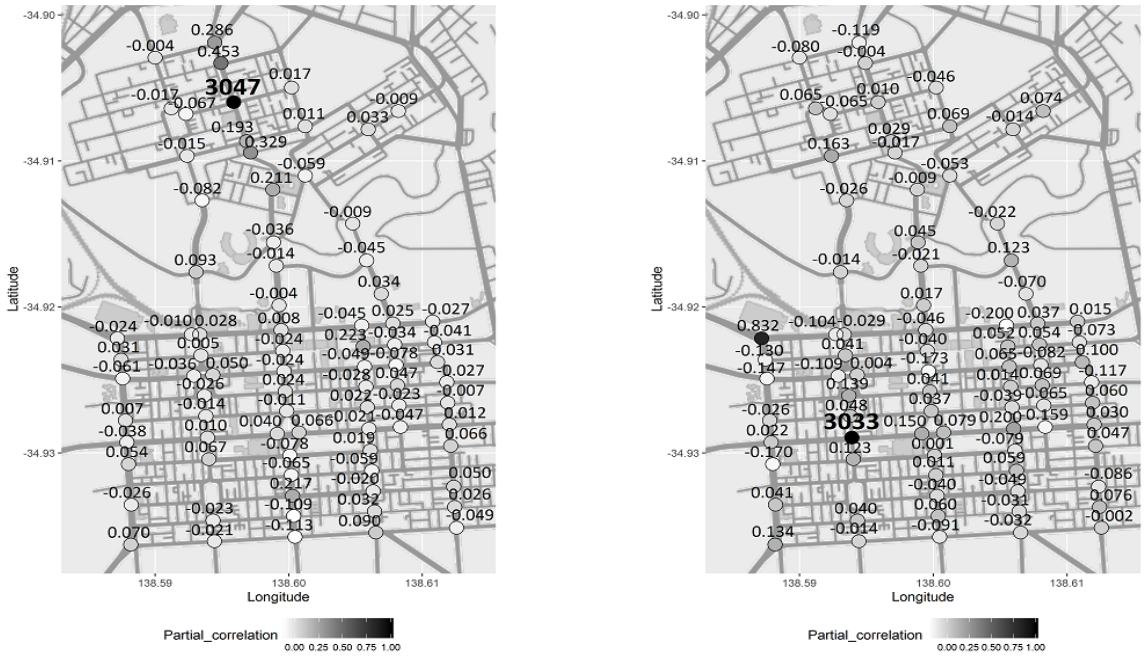
Figure 2 demonstrates the partial correlations of traffic volumes between two sites and others in the city centre of Adelaide, Australia. In calculating these partial correlations, a R package named "ppcor" is used. A more detailed description of the calculation procedure can be found in [26]. In the figure, the correlation strength is indicated by the degree of darkness inside the circles. The left panel of the figure is for site 3047, and the right panel is for site 3033. The left panel clearly shows that at site 3047, the traffic volume is mainly correlated with the intersections nearby vertically but not horizontally. This is consistent with the conventional wisdom that traffic volume at a location is usually correlated with that at adjacent locations downstream or upstream. However, the right panel of Figure 2 tells a completely different story, which shows that the volume at site 3033 is strongly correlated with that at a site in the upper left corner of a loop. Clearly in this case, it is not the adjacent locations that can provide the most relevant information for forecasting the volume at site 3033. In other words, only considering neighbouring locations downstream or upstream in a model (as the common practice in the literature, e.g., [11, 27, 28]) may lead to losses of valuable information, and consequently suboptimal design of the forecasting framework.

To make the forecasting task even more challenging, Figure 3 shows the correlation strengths between site 3047 and other sites at two different hours of a day (1 a.m. and 6 a.m.). The left panel shows that the traffic volume at 1 a.m. is generally not correlated with that at any of the sites located in the lower half of the network. However, the same cannot be said for the volume at 6 a.m. (right panel). At this time, the volumes at two locations in the lower half of the network become notably correlated with the volume at site 3047. A clear evidence for the existence of a similar spatial-temporal pattern was also found in speed [21].

Moving to another aspect of traffic data. It is reasonable to expect that the characteristics of traffic volume to be different at different levels of aggregation. The aggregation considered here can be either spatial or temporal, or both. In the literature, the aggregation in the temporal dimension has attracted relatively more attention, and it is widely accepted that the level of aggregation in the temporal dimension can have an impact on the forecasting accuracy. More specifically, an aggregation based on a longer time window normally results in a smoother series, thus facilitates pattern identification on a longer time horizon. But this comes at the expense of loosing useful information at higher frequencies and thus possible deterioration of the forecasting accuracy. Consequently, it is important to keep a good balance when selecting the time window [29, 30, 31].

Similarly, aggregations can also be done in the spatial dimension to provide a potential avenue for improving the forecasting accuracy. Intuitively speaking, a series aggregated in the spatial dimension for a larger area is expected to be smoother, thus facilitate the identification of the general trend of traffic volume for the whole area. Whereas a single location series is more noisy but contains location-specific characteristics which may be useful for forecasting traffic volume at a specific location. Having the information been identified separately at different spatial aggregation levels, the challenge is thus to devise a forecasting framework which can produce more accurate forecasts by optimally combining the identified information.

To summarise, three challenges in short-term traffic flow forecasting has been identified in relation to the features discussed in this section: *First*, how to effectively model the multiple seasonal patterns as shown in Figure 1? *Second*, how to improve forecasting accuracy by taking advantage of the spatial correlations as demonstrated in Figure 2? Specifically, how to not just consider adjacent locations, but choose the most relevant locations from a large number of candidates in a broad area? Moreover, how to accommodate the dynamic interactions between the temporal and spatial patterns as shown in Figure 3, so not only the information both at



(a) Site 3047

(b) Site 3033

Fig. 2: The layout of the 81 chosen intersections in the city centre of Adelaide, Australia. The 81 intersections are shown as circles with the partial correlations over the entire period for sites 3047 (left panel) and 3033 (right panel) indicated both by the degree of darkness inside the circles the numbers above.

(a) Site 3047, 1 a.m.

(b) Site 3047, 6 a.m.

Fig. 3: The partial correlations of the traffic volumes of site 3047 at 1 a.m. (left panel) and 6 a.m. (right panel), respectively. The 81 chosen intersections in the city centre of Adelaide, Australia are shown as circles with the partial correlations over the entire sample period indicated both by the degree of darkness inside the circles the numbers above.

the most relevant locations but also at the most relevant time is considered. *Finally*, how to improve the forecasting performance by taking advantage of the smoother dynamics of the spatially aggregated series while at the same time making use of location-specific information at single locations?

III. Forecasting methodology

A combined forecasting framework is proposed to accommodate the main characteristics of traffic volume identified in Section II. Briefly, a gradient boosting algorithm is used to deal with the first two characteristics regarding the seasonality, the spatial pattern and interactions between the temporal and spatial patterns; for the third, a hierarchical forecast reconciliation procedure is used for optimally combining the forecasts produced at different spatial aggregation levels.

A. Gradient boosting

The characteristics of traffic flow identified in Section II suggest that a good forecasting model of traffic volume should not only take account of the multiple seasonal and spatial patterns, but do so dynamically depending on the hours of a day. To enable this flexibility, the component-wise gradient boosting procedure (CWGB) that originated from the machine learning literature is used [22, 32, 33, 34].

In the gradient boosting framework, given a training dataset $D = \{(\boldsymbol{x}_t, y_t) : t = 1, \ldots, T\}$ of size T and a loss function L, such as the sum of squared errors, $L_2 = \sum_{t=1}^{T} (y_t - m(\boldsymbol{x}_t))^2$ with $m(\boldsymbol{x}_t)$ denoting the conditional mean of y_t given $\boldsymbol{x}_t$ at time t. The component-wise L_2 gradient boosting proceeds as follows:

(1) Initialise the procedure by setting $F_0 = m$, where m is the unconditional mean of $\boldsymbol{y} = (y_1, \ldots, y_T)$ and F_0 is the forecasting model at the initial iteration. Define a set of base functions, $f_n(x)$, with pre-determined forms, index n and argument x chosen from a set of observed covariates. For example, the set of base functions could be the set of all the linear functions in one argument, or polynomial functions in one or more arguments, etc.

(2) Increase the iteration i by 1, and calculate the negative gradient $\boldsymbol{g}_i$ evaluated based on the model obtained at the previous step, F_{i-1}. That is, for each t,

$$g_{t,i} = -\frac{\partial L(y_t, F_{i-1}(\boldsymbol{x}_t))}{\partial F_{i-1}(\boldsymbol{x}_t)}.$$

In the case of L_2 boosting, the negative gradient can be replaced by the residuals obtained from regressing $\boldsymbol{F}_{i-1} = (F_{i-1}(\boldsymbol{x}_1), \ldots, F_{i-1}(\boldsymbol{x}_T))$ on $\boldsymbol{y}$.

(3) Carry out an exhaustive search by fitting each base function separately using the negative gradient obtained at the previous step as the target variable, and record the sum of squared residuals obtained from all the individual regressions.

(4) Choose the base function, denoted by b_i that has the minimum sum of squared residuals recorded at the previous step. Update the function estimate at iteration i by setting

$$F_i = F_{i-1} + \nu b_i,$$

where ν is a predefined weight with the value belongs to $(0, 1]$.

(5) Stop the algorithm if the number of iteration has reached a pre-determined value, otherwise go to step 2.

As indicated in the procedure above, at each iteration, the boosting algorithm attempts to drive down the negative gradient evaluated at the current function estimate in the direction of one base function that is the closest to the negative gradient in terms of the Euclidean distance, hence the name gradient boosting or functional gradient decent. Also at each iteration, the function estimate is updated by combining the existing one with one extra base function (usually, a function in one argument chosen from the set of covariates), thus the term component-wise.

One important feature of CWGB that is particularly beneficial to the present study is that at each iteration, only the base function with one specific covariate that is most useful in driving down the negative gradient at the current function estimate is chosen to be included. In other words, at the end of iterations, all the base functions combined will be just those which are most relevant in explaining the variation in the target variable. Consequently, when applying the algorithm to traffic data in this study, only those covariates representing the temporal and spatial locations that are most relevant to the temporal-spatial patterns shown in Section II will be included in the final forecasting model. Moreover, the way of choosing lags and locations in CWGB is based on the statistical evidence presented in the data measured by a loss function instead of relying on subjective judgements or some arbitrary significance levels, and is generally applicable to the forecasting of volume at any location in any time period without requiring any prior knowledge on the characteristics of traffic flow at specific locations.

B. Model specification

The boosting procedure introduced above explicitly assumes the final output model to be a linear combination of the base functions chosen at all the iterations. Consequently, the problem of model specification becomes choosing the functional form of the base functions and the type of covariates to be used as the input arguments. Essentially, after the algorithm is terminated, the final output model is a special case of the generalised additive model [35].

For simplicity, we restrict our attention to linear base functions in one input argument only. In terms of the choice on input arguments, since the boosting procedure itself can select both the appropriate lag length and locations from a given set of candidates, the historical observations up to a maximum lag length at all locations can be included. There is one more point on the model specification, which is concerned with the spatial-temporal interactions as shown in Figure 3, i.e., depending on the hour of a day, the correlation between locations can vary significantly. To accommodate such interactions in the forecasting model, the dataset is divided into 24 hourly subsets. Consequently, the boosting procedure is used separately on each of the 24 subsets.

In summary, the final forecasting model constructed by the boosting algorithm takes the general form:

$$y_{h,d,l} = m_{h,l} + \nu \sum_{i=0}^{M} (\beta_{1,h,l,i} + \beta_{2,h,l,i} x_{h,d,l,i}) + \varepsilon_{h,d,l},$$

where the traffic volume at hour h, day d and location l is defined to be a linear combination of the unconditional mean of the traffic volume $m_{h,l}$, a residual term $\varepsilon_{h,d,l}$, and all the linear base functions selected in the boosting procedure, M is the maximum number of iterations and ν is a weight. The parameters $\beta_{1,h,l,i}$ and $\beta_{2,h,l,i} x_i$ are estimated in each iteration of the boosting procedure for specifying the linear base function with input argument $x_{h,d,l,i}$. Based on this specification, it is clear to see that CWGB fits into the general framework of boosting [34], and can be described as a boosting procedure with linear base functions and L_2 loss (as opposed to trees and exponential loss). Moreover, it is a form of ensemble learning, in which each chosen base function represents a simple linear model with only one covariate[36].

For this study, $M = 1,000$ and $\nu = 0.3$. In our experiments, the forecasts produced by an alternative with maximum iterations being determined by cross validation are less accurate. This phenomenon is similar to that observed and explained in [34] for adaptive boosting.

C. Hierarchical reconciliation of forecasts

For a time series at any spatial aggregation level, the boosting procedure described above can be used to generate forecasts at that level. For example, at the most disaggregated level, the forecasts of traffic volumes at each location can be generated using all traffic volumes at different locations in the network; at the most aggregated level, the forecasts of traffic volumes at the network level can be produced by simply using the sum of volumes at all the locations in the network. Moreover, the locations in the network can be grouped into different subsets, and for each subset, the aggregated series and the corresponding forecasts can be obtained. Due to these different ways of grouping, a hierarchical structure for the forecasts across different aggregation levels emerges. That is, the sum of forecasts at the locations that belong to a group should be equal to the forecasts of the aggregated volume for that group as a whole. This structure of traffic volume forecasts in the spatial dimension naturally fits into the hierarchical reconciliation approach (HR) of [23].

HR is a regression based forecast combination approach, which uses weighted least-square regressions in combination with parameter restrictions to allow a reconciliation of the forecasts across different aggregation levels. The motivation of [23] is to produce forecasts that are consistent at different aggregation levels, and thus provide consistent information for the decision making at all levels. In the present case of traffic volume forecasting, we are primarily interested in the forecasts at each single location instead of the aggregated volume forecasts of the whole area. The main reason of using this approach in traffic volume forecasting is that it allows the forecasts generated at different aggregation levels to be combined in a natural way so the information that is better identified at different aggregation levels can be used for the forecasting of volumes at any location. For example, at the most aggregated level, the general trend that can be better identified using the aggregated series with less location-specific noise, thus providing additional information for inferring or adjusting the volume forecast at any specific single location, whereas the forecasts generated based on the most disaggregated series rely more on location-specific information, which can be used to infer or adjust the volume forecasts at more aggregated levels or at other single locations.

In the case of the spatial reconciliation of traffic volume forecasts, the reconciled forecasts at each time point can be generated using the following steps:

(1) Stack individual forecasts at all aggregation levels as a column vector,

$$\boldsymbol{y}_t = \left(\boldsymbol{y}_t^{1\prime}, \ldots, \boldsymbol{y}_t^{I\prime}\right)',$$

where I is the total number of aggregation levels. At each aggregation level i ($i = 1, \ldots, I$),

$$\boldsymbol{y}_t^i = \left(y_{t\,i\,1}, \ldots, y_{t\,i\,n_i}\right)'.$$

That is, at each aggregation level i, the corresponding vector $\boldsymbol{y}_t^i$ contains all the individual forecasts at that aggregation level. The total number of individual forecasts at level i is denoted by n_i.

(2) Create a matrix $\boldsymbol{S}$ of size $M \times n_1$ with $M = \sum_{i=1}^{I} n_i$. For each row m of $\boldsymbol{S}$, set the entries corresponding to the locations (at the most disaggregated level) contained in the spatial area represented by row m, to 1 and 0 otherwise. For example, the matrix $\boldsymbol{S}$ for a total number of three individual locations at the most disaggregated level with $\boldsymbol{y}_t^1 = (y_{t\,1\,1}, y_{t\,1\,2}, y_{t\,1\,3})'$ and three aggregation levels,

$$\begin{aligned}\boldsymbol{y}_t =& \left(\boldsymbol{y}_t^{1\prime}, \boldsymbol{y}_t^{2\prime}, \boldsymbol{y}_t^{3\prime}\right)' \\ =& \left((y_{t\,1\,1}, y_{t\,1\,2}, y_{t\,1\,3}), \left(\sum_{l\in\{1,2\}} y_{t\,1\,l}, \sum_{l\in\{1,3\}} y_{t\,1\,l}, \sum_{l\in\{2,3\}} y_{t\,1\,l}\right), \left(\sum_{l\in\{1,2,3\}} y_{t\,1\,l}\right)\right),\end{aligned}$$

is defined as

$$\boldsymbol{S} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}.$$

(3) Construct a diagonal weighting matrix W_t of size M $\times$ M with the diagonal element at row r ($r = 1, \ldots M$), equal to the sample variance of the residuals obtained from the estimation stage of the model for forecasting the corresponding rth element in $\boldsymbol{y}_t$. For other choices of the weighting matrix, see [23].

(4) Finally, use the weighted least square procedure to fit the model

$$\boldsymbol{y}_t = S\boldsymbol{\beta}_t + \boldsymbol{\varepsilon}_t$$

with

$$\boldsymbol{\beta}_t = (\boldsymbol{S}'\boldsymbol{W}_t\boldsymbol{S})^{-1}\boldsymbol{S}'\boldsymbol{W}_t\boldsymbol{y}_t,$$

where $\boldsymbol{\varepsilon}_t$ can be thought as the reconciliation errors for all the locations and at all aggregation levels. The obtained $\boldsymbol{\beta}_t$ is a vector of length n_1 and contains the reconciled forecasts for all the locations at the most disaggregated level at time t.

In summary, the proposed forecasting framework, denoted as CWGB-HR hereafter, consists of two steps. *Step I*: all the individual forecasts at different aggregation levels are produced by using the boosting procedure described in Subsection III-A; and *Step II*: the hierarchical reconciliation introduced above is then used to reconcile all the individual forecasts and produce $\boldsymbol{\beta}_t$, the reconciled forecasts for all the individual locations at the most disaggregated level.

IV. A CASE STUDY

To demonstrate the implementation of CWGB-HR, and particularly the efficacy of each component, the same hourly volume data collected from loop detectors at intersections of the city centre of Adelaide is used in this section. A three-year moving window is used for model estimation. The three-year moving window estimation is done by firstly dividing the whole forecasting period into blocks. For each block, the preceding three years of data is used for model estimation and producing the forecasts in that block. Out of sample one-hour-ahead forecast is produced for the period from 2013-01-04 to 2014-06-19.

The input arguments of the base functions are chosen to be one to eight hours lagged, one day lagged and one week lagged volumes at all intersections. In addition, a refinement is made by including the interaction of one-day lagged volumes with day-of-week indicator variables. In effect, this refinement creates six more variables for each location. The reason for this refinement is due to the strong weekly pattern of the traffic volume shown in Figure 1 and the significant improvement made by this refinement in electricity load forecasting [37, 38], in which similar seasonal patterns exist. More specifically, letting $y_{h,d,l}$ denotes the traffic volume at hour h, day d and location l, and $\mathbf{1}_{d,l,w}$ $w = 1, \ldots, 6$, for the day-of-week indicator variables (from Monday to Saturday, respectively). Then the set of input arguments used is,

$$\bigcup_{l=1}^{81} \{\{y_{h-j,d,l} : j = 1, \ldots, 8\} \cup \{y_{h,d-1,l}, y_{h,d-7,l}\} \cup \{\mathbf{1}_{d,l,w} y_{h,d-1,l} : w = 1 \ldots 6\}\}. \quad (1)$$

A total number of 1,296 input arguments are considered, and each one corresponds to a unique base function.

Figure 4 compares the averaged hourly residuals of a week obtained by using CWGB-HR on the full moving window dataset from 2010-01-01 to 2013-01-03 (upper panel), and 24 subsets in the same sample period corresponding to the hours of a day (lower panel). It can be seen that there is a clear daily pattern in the residuals obtained based on the full dataset. In contrast, the daily pattern is largely alleviated in the lower panel, when the 24 hours are treated separately. This is due to the fact that when treating the forecasting problem separately depending on the hours of a day, the intra-day pattern is held relatively constant in any one of the subsets. Consequently, each individual model does not need to take into account of the intra-day pattern, and the intra-day pattern is captured instead by the different parameter estimates across the individual models produced for the 24 subsets.

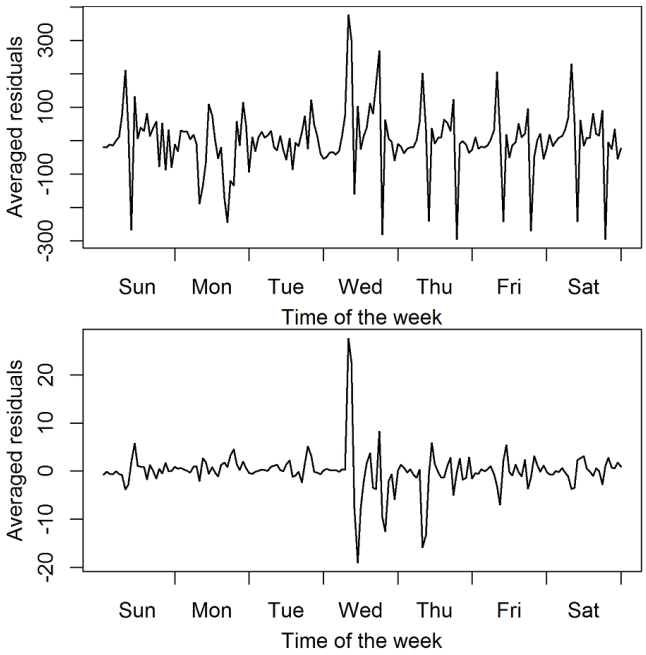
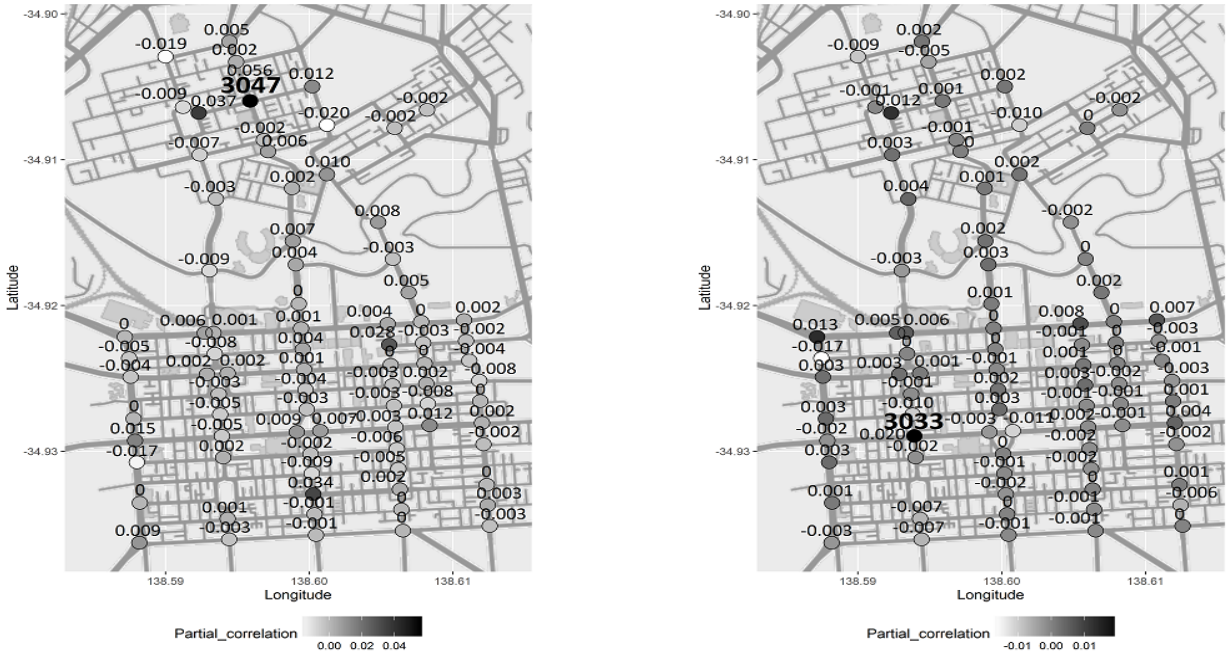


Fig. 4: Averaged residuals for the hours of a week based on the first moving window dataset from 2010-01-01 to 2013-01-03 at site 3047. The upper panel is based on the residuals obtained by using CWGB-HR on the full moving window dataset. The lower panel is based on the residuals obtained by using CWGB-HR separately on 24 subsets of the data representing the 24 different hours of a day. The information shown is unnormalised, thus having the unit of vehicle count.

In regard to the modelling of the spatial pattern, Figure 5 plots the sum of parameters in the forecasting models for each hour of a day at sites 3047 and 3033, respectively. It can be seen that in both panels of Figure 5, the relative magnitude of the estimated parameters shown by the degree of darkness inside the circles (and the numbers above the circles) roughly reflects a similar pattern of the partial correlations observed in Figure 2. However, a noticeable difference in the pattern shown in Figure 5 from that in Figure 2 is that more weights are given to some intersections which are distant to site 3047 (left panel) or 3033 (right panel). This difference further confirms that the spatial pattern in traffic flow is not restricted to between adjacent locations. Moreover, the difference observed also does not conflict with the early observation in Figure 2, since the quantity used in calculating the partial correlations in Figure 2 is the volumes in the same time interval, whereas in Figure 5, hourly, daily and weekly lagged volumes are used instead.

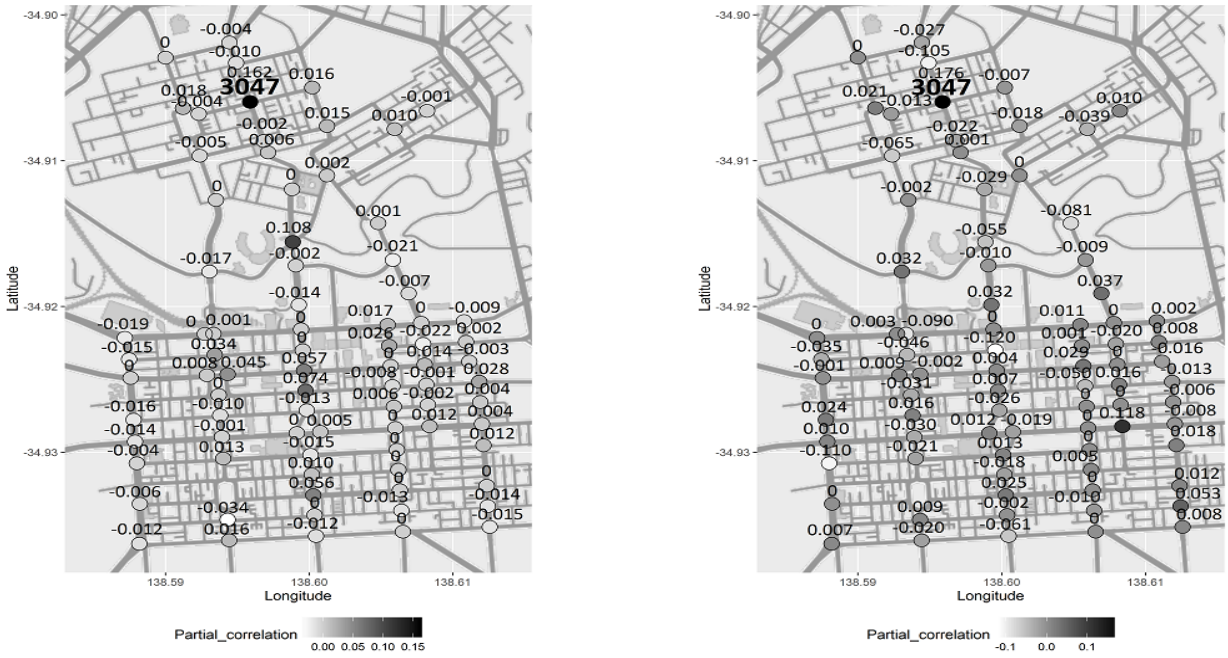
Similarly, the sum of parameters in the forecasting model of site 3047 at 1 a.m. and 6 a.m. shown in Figure 6 also indicates that the spatial pattern changes depending on the time of a day, i.e. interactions between patterns in the temporal and spatial dimensions.



(a) Site 3047

(b) Site 3033

Fig. 5: Sum of parameters in the forecasting models of sites 3047 (left panel) and 3033 (right panel) for each site and each hour of a day, obtained based on the first moving window dataset.



(a) Site 3047, 1 a.m.

(b) Site 3047, 6 a.m.

Fig. 6: Sum of parameters in the forecasting models of site 3047 for each site and at 1 a.m. (left panel) and 6 a.m. (right panel), respectively, obtained based on the first moving window dataset.

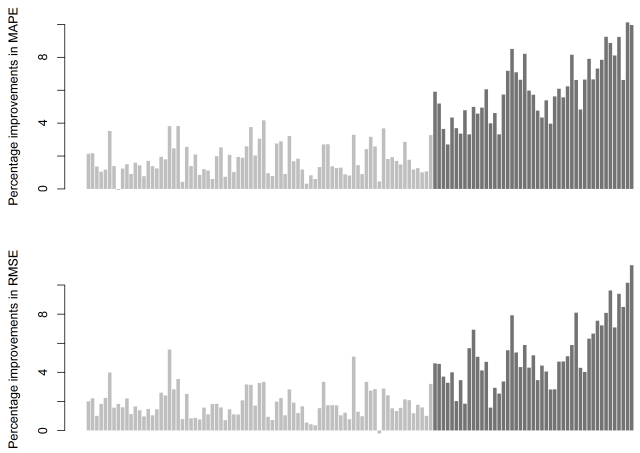


Fig. 7: The percentage improvements in MAPEs/RMSEs of the hierarchically reconciled forecasts over individual forecasts based on an one-hour-ahead forecasting horizon for the period form 2013-01-04 to 2014-06-19. The 81 individual sites are shown in grey colour, and the spatially aggregated sites are shown in black colour with the level of spatial aggregation increasing from the left to the right.

Lastly, Figure 7 demonstrates the efficacy of HR described in Subsection III-C. This figure shows the percentage improvements in the accuracies of one-hour-ahead forecasts obtained by using CWGB-HR over that from using CWGB only. In terms of the Mean Absolute Percentage Error (MAPE), it is clear that the improvements obtained by using HR are positive for almost all the sites. Although the magnitudes are small for a number of individual sites (in grey colour), the fact that they are positive across almost all the sites indicates strongly the advantage of HR. Furthermore, the improvements become larger as moving from individual sites to spatially aggregated sites (in black colour), and reach the highest of over ten percent for the forecasts of the highest aggregation level at the farthest to the right side. The same conclusion can also be drawn from the improvements measured by Root Mean Squared Error (RMSE).

The consistent improvements made by using HR should come with no surprise. To explain, the difference in the variations of the volumes at different spatial aggregation levels suggests that the aggregated volume normally has less variation. Intuitively, for most travellers, they may need to travel to a destination regularly, but the specific route they choose may depend on various conditions and is highly variable. This is especially the case for traffic flows in an urban area where many alternative routes are available. By aggregating the volumes, the general trend of the traffic flow in that area can be better identified. Consequently, the forecasts of aggregated volumes put more emphasis on the general traffic dynamic of the area as a whole. Whereas the forecasts for individual sites are more noisy, but contain location-specific information that is averaged out in the aggregated series. The disaggregated forecasts therefore put more emphasis on location-specific features of the traffic flow that is useful for forecasting location-specific volumes. Combining the strength of both, HR allows the different information to be optimally combined and produces the final forecasts based on the better identified information at all aggregation levels. This combination consequently leads to the consistent improvement in the forecasting accuracy as shown in Figure 7.

Overall, this section has demonstrated that the main features of traffic flow identified in Section II can be accommodated satisfactorily by using CWGB-HR. Two major advantages, namely, the flexibility and autonomy in learning and filtering useful and rich information in both the temporal and spatial dimensions also emerge from this case study. In order to further demonstrate the performance of the proposed framework, its forecasting accuracy is compared with that of three mainstream methods in the next section.

V. Performance comparison

Three common approaches of traffic forecasting, namely Seasonal Autoregressive Integrated Moving Average (SARIMA), Kalman filter model and random forest are used in this section for comparison purposes. For the SARIMA and Kalman filter model, they can be both expressed in a state-space form. Thus, the two benchmarks used are actually state-space models of different specifications. In terms of the SARIMA, a SARIMA$(1,0,1)(0,1,1)_S$ with weekly seasonality S, as recommended by [39] is used. The Kalman filter model is implemented using the function "fitSSM" in the R package, "KFAS" (see, https://cran.r-project.org/web/packages/KFAS). A more detailed description of the specifications of both SARIMA and Kalman filter model can be found in the Appendix. In regard to the random forest benchmark, the results are obtained using the R package "randomForest" with the tuning parameters being determined by cross validation (see,https://cran.r-project.org/web/packages/randomForest). Similar to the implementation of CWGB-HR, all three benchmarks are specified separately for each hour of a day unless indicated otherwise. In addition, the covariates included in these models differ across the datasets, and described later case by case. A more detailed description of the benchmark models and the covariates used in each case can be found in the Appendix.

To make this comparison comprehensive, three datasets from different countries, and with different time resolutions and different levels of network complexity are used for forecast evaluations at three different forecasting horizons:

Dataset I: The hourly traffic volumes collected from loop detectors at 81 intersections in the city centre of Adelaide, Australia for the period from 2010-01-01 to 2014-06-19. The study site is shown in Figure 2. This dataset is the same as the one used in Sections II and IV.

Dataset II: The ten-minute traffic volumes collected from loop detectors at 92 stations in a road network of the twin cities of Minnesota, USA (Figure 8), for the period from 2010-6-3 to 2010-11-20. In comparison with Dataset I, the volume in this dataset was not collected at intersections, and each circle in Figure 8 represents two stations measuring traffic flow in two opposite directions, respectively.

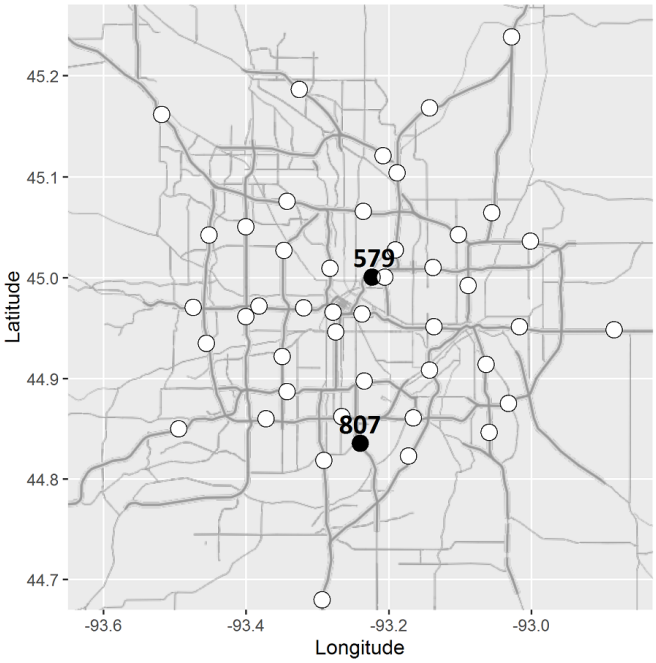


Fig. 8: The layout of the 92 chosen stations in the arterial network of the twin cities of Minnesota, USA.

Dataset III: The three-minute traffic volumes collected from seven loop detectors for the period from 2000-4-1 to 2000-4-30 on an arterial road of Athens, Greece (Figure 9). In which, the seven loop detectors are located at four locations with three of the locations having two loop detectors measuring flow in two opposite directions. This dataset is the same as the one used in [40] and [15]. One feature of the dataset that is immediately apparent from Figure 9, is the significantly reduced network complexity compared with that of the previous two datasets.

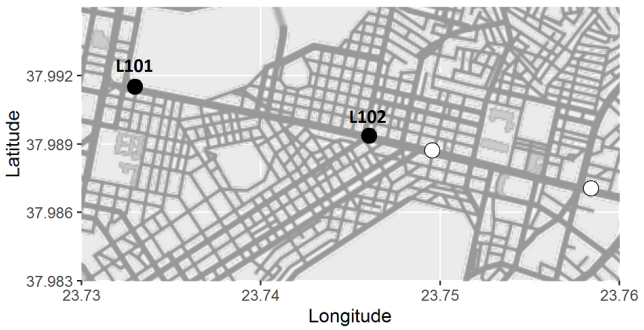


Fig. 9: The layout of the seven loop detectors contained in the traffic volume dataset for an artery of Athens, Greece.

Regarding the data sources, all three datasets were collected by loop detectors. As widely acknowledged in the traffic flow community, loop detector data can be noisy due to various reasons, e.g., loop detector malfunction, vehicle double counting, communication issues, and etc. [41]. However, currently loop detector data are still the most widely used data type in Traffic Engineering, and it is likely to remain so in the foreseeable future (perhaps till the wide adoption of connected vehicles). Largely for this reason, in the traffic forecasting literature, loop detector data are commonly used as the ground truth to test different forecasting methods' performance [e.g. 42, 40, 15]. This is also the case in the current study. A major difference between this study and many existing studies is that we use three distinctive datasets, and each dataset contains data collected at many different locations at different time periods, which diminishes the potential risk caused by unreliable measurements.

In measuring the forecasting accuracies, two measures are used: RMSE and a modified MAPE, denoted as MAPE_L with L indicating the level of the observed volume below which the corresponding Absolute Percentage Error (APE) has zero weight in the averaging process. They are defined as:

$$\text{RMSE} = \left(\frac{\sum_{t=1}^{T} (y_t - \bar{y}_t)^2}{T} \right)^{1/2},$$

and

$$\text{MAPE}_L = \frac{\sum_{t=1}^{T} \left(\mathbf{1}_{\{y_t > L\}} |y_t - \bar{y}_t| \right) / y_t}{\sum_{t=1}^{T} \mathbf{1}_{\{y_t > L\}}},$$

respectively; where $\bar{y}_t$ is the forecast for observation t, and $\mathbf{1}_{\{y_t > L\}}$ is an indicator function which takes value one if the tth observed volume, y_t is greater than L, and zero otherwise.

While the use of RMSE is common, MAPE_L is different from the usual MAPE used in the literature. This choice of the accuracy measure is motivated by the nature of the problem in this study. Specifically, the magnitude of a chosen accuracy measure for a forecasting problem should truthfully reflect the loss incurred in terms of the decision maker's attitude toward risk, and thus provide correct guidance on assessing the performance of competing models. In this study , RMSE gives an averaged magnitude of how close the forecasted volumes are to the observed with more weight given to large deviations in the averaging process. Thus it can be considered as a useful measure especially when large forecast errors are considered as less desirable. However, the same can not be said to the usual MAPE measure. Consider the case of using MAPE for measuring the accuracy of volume forecasts in early morning, for example from 3:00 to 4:00 a.m. The observed volume in this time interval is expected to be very low or at zero. Thus one unit deviation of the forecasted volume from the observed would lead to a very large APE. The resulting MAPE will therefore, be inflated. Obviously, this inflation in MAPE mismatches the true risk faced in traffic volume forecasting, where the accuracy of a forecasting model should be judged mainly based on the accuracy in peak periods when the future volume information is needed most. And a model selected based on MAPE measure has the tendency to perform better at forecasting volumes in off-peak intervals. In contrast, the use of MAPE_L alleviates this issue by assigning more weight to the forecasting accuracy for peak hours.

A. *One-hour-ahead forecasting comparison*

The hourly traffic volumes in Dataset I (Figure 2) is used in this comparison. The forecasts are produced using a three-year moving window procedure with the forecasting period chosen to be from 2013-01-04 to 2014-06-19. The one-hour-ahead forecasts of sites 3033 and 3047 are chosen from the

81 sites for demonstration purposes. For this comparison, the input arguments used in CWGB-HR and random forest are identical to that shown in Equation (1), whereas the covariates for SARIMA and Kalman filter models are restricted to those at local locations only. Similar to the implementation of CWGB-HR, all the benchmark models are estimated based on subsets of data for the hours of a day. The obtained accuracies are reported in Table I. It is clear that the forecasting performance of CWGB-HR is superior in terms of both RMSE and MAPE_{100}.

TABLE I: A comparison of overall MAPE_{100} and RMSE of one-hour-ahead forecasts of traffic volume for sites 3033 and 3047 in the city centre of Adelaide, Australia.

	MAPE_{100} (%)		RMSE	
	Site 3033	Site 3047	Site 3033	Site 3047
SARIMA	10.55	12.70	366.93	213.20
Kalman filter	10.36	12.12	269.99	208.48
Random forest	7.79	9.43	266.81	165.46
CWGB-HR	5.52	6.72	141.36	101.31
Observations	12747	12551	12768	

B. Ten-minute-ahead forecasting comparison

Dataset II (Figure 8) is used in this comparison. The forecasting period spans from 2010-10-31 to 2010-11-20, and all the preceding data (from 2010-6-3 to 2010-10-30) is used for model estimation. Since the forecasting horizon has now been shortened significantly, the most recently observed volumes should have a more significant role in inferring the future ones, and the observed volumes in distant past such as one week lagged volume are expected to have minimum influence on the forecasts. Consequently, the set of input arguments used in this comparison is chosen to be

$$\bigcup_{l=1}^{92} \big\{\{y_{h-j,d,l} : j = 1, \ldots, 9\} \cup \{y_{h,d-1,l}\}\big\},$$

for CWGB-HR and random forest. Similarly, the covariates included in the SARIMA and Kalman filter models are restricted to those at local locations only. Furthermore, all the benchmark models and CWGB-HR are estimated based on subsets of data for the hours of a day. The obtained ten-minute-ahead forecasting accuracies for stations 579 and 807 are reported in Table II.

TABLE II: A comparison of overall MAPE_{10} and RMSE of the ten-minute-ahead forecasts for stations 579 and 807 in the arterial network of the twin cities of Minnesota, USA.

	MAPE_{10} (%)		RMSE	
	Station 579	Station 807	Station 579	Station 807
SARIMA	10.25	10.19	30.78	32.14
Kalman filter	16.89	16.77	44.92	48.15
Random forest	9.19	9.22	27.15	28.43
CWGB-HR	8.78	8.81	24.14	26.28
Observations	3017	3005	3024	

Overall, a similar result compared to that from the Adelaide dataset is obtained. CWGB-HR has achieved a significant gain in the forecasting accuracies.

C. Three-minute-ahead forecasting comparison

Dataset III (Figure 9) is used in this comparison. The forecasting period is chosen to be from 2000-4-16 to 2000-4-30, and the model estimation is based on all the preceding data (from 2000-4-1 to 2000-4-15). For the same reason as in the previous comparison, the set of input arguments used in this case is modified to be

$$\bigcup_{l=1}^{7} \{y_{h-j,d,l} : j = 1, \ldots, 20\}, \tag{2}$$

for CWGB-HR and random forest, and the local location only subset of (2) for SARIMA and Kalman filter models. Due to the limited time span of the data, the SARIMA implemented for this dataset does not included one week lagged error term. In other words, a SARIMA(1,0,1)(0,1,0)$_S$ is used instead. Following the same consideration for the intra-day dynamic as described in [40, 43], the six day periods are used instead to define the subsets of data for model estimation and forecasting.

TABLE III: A comparison of overall MAPE_{10} and RMSE of the three-minute-ahead forecasts for loop detectors L101 and L102 on an artery of Athens, Greece.

	MAPE_{10} (%)		RMSE	
	L101	L102	L101	L102
SARIMA	26.13	26.66	12.24	12.70
Kalman filter	24.68	25.87	50.48	40.55
Random forest	33.51	22.64	41.35	23.05
CWGB-HR	16.96	16.45	15.37	18.37
Observations	7098	7107	7143	

The obtained accuracies for two locations (L101 and L102) based on the three-minute-ahead forecasts from 2000-4-16 to 2000-4-30 are shown in Table III. Interestingly, the accuracy obtained by SARIMA model is now better than that of CWGB-HR in terms of RMSE, whereas the performance of random forest becomes unstable.

D. Discussion

Taken the results of the three comparisons together, a clear pattern concerning the performance of CWGB-HR in relation to the characteristics of the datasets emerges. For Dataset I, the volumes are collected from a complex city centre network at a one-hour time resolution. In this case, both the information at a longer temporal horizon and at a wider spatial horizon can have significant roles in inferring future volumes at any given location. Consequently, lag selections may significantly affect the forecasting accuracy of a model, i.e., the lag selection for the volumes on the same day but at different hours, the lag selection for the volumes at the same hour but on different days, and the selection of these lagged volumes at different locations. When CWGB-HR is used, these lag selections can be done automatically and efficiently by filtering through a large quantity of information in the dataset, which enables the final model to take advantage of the most useful information and thus achieve a forecasting performance significantly better

than all three benchmarks (over 40% better than its closest competitor).

For Dataset II, while the spatial complexity of the network is similar to that of Dataset I, the complexity of the forecasting problem in the temporal dimension has been reduced due to the shorter observation interval (i.e., ten minutes). At this shorter time interval, the future volume becomes more correlated with the most recently observed volume, which implies that the forecasting accuracy is less affected by the way how other distantly lagged volumes are selected. This is consistent with the accuracy reported in Table 2, where the improvement of the forecasting accuracies made by using CWGB-HR is around 5 to 20% in comparison with the top competitors.

The reduction in the improvement of the forecasting accuracy made by using CWGB-HR is even more pronounced in Dataset III, in which the complexity of the forecasting problem is further reduced in both temporal and spatial dimensions (a shorter observation interval and a simpler network). In this case, the performance of CWGB-HR, as shown in Table 3, even becomes slightly worse than that of SARIMA in terms of RMSE. A further investigation of the estimated parameters by CWGB-HR reveals that at a three-minute interval, one period lagged volumes have a dominant role in terms of the magnitude of the estimated parameters, and therefore in inferring the future volumes. Since this lag length is covered by SARIMA, a similar forecasting performance should be expected. In addition, for a network as simple as being a single straight road of around two and a half kilometres, the traffic conditions at different locations should be highly correlated. Thus less additional information can be uncovered by broadening the forecasting model's spatial coverage.

Overall, CWGB-HR offers a highly flexible and automated approach to learn useful information in both the temporal and spatial dimensions, which improves the forecasting accuracy particularly for cases with complex road networks and long forecasting horizons.

The last point which deserves further comment is on the choice of the competing models used in this section. One might argue that the multivariate version of the competing models should be the more appropriate alternatives to compare with. Moreover, the covariates used in two of the competing models (SARIMA and Kalman filter) are restricted to the volumes at local locations only while CWGB-HR and random forest have access to the volume information at all locations. It seems that it would only be fair if all the models have access to the same amount of information. In fact, these concerns further illuminate the advantage of CWGB-HR both from the computation and forecasting accuracy point of view.

Consider the forecasting problem in Dataset III first. Due to the high observation frequency and the simplicity of the network, a reasonable amount of additional parameters can be specified in the multivariate version of both the competing models (SARIMA and Kalman filter), and the computation burden may be handled satisfactorily. However, the same cannot be done when moving to Dataset I and II. In these cases, the number of parameters would increase significantly due to the longer lags for each locations and at nearly 100 locations. This increase in the number of parameters not only means that the estimation of the parameters in the two competing models (especially in Kalman filter model) becomes intractable, but more importantly, it can lead to a deterioration in the forecasting accuracy due to the introduction of many insignificant predictors [15]. Admittedly, this deterioration can be alleviated to a certain degree by a careful selection of the covariates. However, this again can only be done at a very high computational cost. Take Dataset I as an example. Because the influence of covariates changes depending on locations and hours of a day, covariates selection would have to be done for each hour of a day and for all 81 locations, i.e., a total number of 1944 (24×81) covariate selections.

Furthermore, when conducting such covariates selections, researchers immediately face the important issue of balancing between two conflicting goals: the simplicity of the model and a better forecasting accuracy. Even worse, the criteria commonly used for covariates selection are not directly related to the goal of achieving higher forecasting accuracy. The insignificance of one parameter at a certain significance level does not necessarily mean that the corresponding covariate does not contribute to achieving a higher forecasting accuracy.

On the other hand, CWGB-HR handles all these issues automatically by the nature of its design. In boosting, the covariates will only enter the final output model if it contributes to driving down the estimation error. Consequently, the resulting model is parsimonious. Moreover, if one accepts the assumption that the distribution for generating future data is the same as the distribution for generating the sampled data, it is shown that the forecasting error is bounded by in-sample estimation error and a small quantity depending on, for example the number of observations in the dataset, etc. [34]. This means that when CWGB-HR is used for forecasting, researchers no longer need to choose a significance level (often chosen arbitrarily) for covariate selections and avoid the risk of deteriorating the forecasting accuracy by having a model that is too simple or has too many insignificant covariates. Overall, CWGB-HR not only offers a computationally tractable and versatile approach to learning useful information from complex datasets, but doing so in an automated way.

VI. Conclusions

Accurate short-term forecast of traffic flow plays an important role in managing modern transportation systems. With the rapid development of data collection and communication technologies, there has been an unprecedented amount of historical data made available for forecasting purposes. A common problem faced by researchers with these large datasets is therefore, how to extract the rich information contained in the datasets and thus provide more accurate forecasts.

In this study, the important characteristics of traffic flow in complex networks is firstly identified using a large dataset collected in the city centre of Adelaide, Australia. These characteristics include the temporal and spatial patterns of traffic volumes, the interactions between the temporal and spatial patterns, and different dynamics of traffic volumes at different spatial aggregation levels. To take advantage of these characteristics, a gradient boosting procedure in combination

with hierarchical reconciliation is proposed for forecasting short-term traffic volume.

The forecasting accuracy of the proposed approach is compared with three competing models (i.e., SARIMA, Kalman filter model and random forest) by using three datasets from different countries, with different network complexity and at three different forecasting horizons. The proposed forecasting framework is found to outperform the three competing models significantly for cases with complex road networks and with a longer forecasting horizon; and perform at least as good as its competitors when the information offered in the datasets for learning is limited. Overall, the gradient boosting based approach offers a powerful and versatile way to learn useful information from large datasets for forecasting purposes, and requires minimum tuning in the implementation stage.

Due to the limited scope, this study only focused on forecasting traffic volumes. However, the flexibility of the gradient boosting based approach demonstrated in this study, means that it can be used for forecasting other quantities related to traffic flow such as travel time, speed, etc. with any other additional information such as weather condition. In particular, congested and uncongested states are not explicitly considered in this paper. Instead, this aspect is accommodated to an extent by the use of multiple individual models for the different time periods. Nevertheless, extending the model to explicitly consider different traffic states by utilizing both volume and speed information is an interesting topic for future research. In the case where information regarding extreme events, such as congestion, road closure, etc. are not available, the corresponding traffic volume can normally be considered as outliers. Then instead of trying to predict these events based on no information, exploring the robustness of the estimated model to these outliers and understanding the effect of these outliers on the forecasts for normal conditions can also be an interesting direction for future research.

ACKNOWLEDGEMENTS

Zuduo Zheng's involvement in this research was partially funded by the Australian Research Council (ARC) through the Discovery Early Career Researcher Award (DECRA; DE160100449).

REFERENCES

[1] E. I. Vlahogianni, J. C. Golias, and M. G. Karlaftis, "Short-term traffic forecasting: Overview of objectives and methods," *Transport Reviews*, vol. 24, pp. 533–557, 2004.

[2] W. Deng, H. Lei, and X. Zhou, "Traffic state estimation and uncertainty quantification based on heterogeneous data sources: A three detector approach," *Transportation Research Part B: Methodological*, vol. 57, pp. 132–157, 2013.

[3] E. I. Vlahogianni, M. G. Karlaftis, and J. C. Golias, "Short-term traffic forecasting: Where we are and where were going," *Transportation Research Part C: Emerging Technologies*, vol. 43, pp. 3–19, 2014.

[4] Z. Zheng and D. Su, "Traffic state estimation through compressed sensing and Markov random field," *Transportation Research Part B: Methodological*, vol. 91, pp. 525 – 554, 2016.

[5] J. McCrea and S. Moutari, "A hybrid macroscopic-based model for traffic flow in road networks ," *European Journal of Operational Research*, vol. 207, pp. 676–684, 2010.

[6] A. Simroth and H. Zahle, "Travel Time Prediction Using Floating Car Data Applied to Logistics Planning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, pp. 243–253, 2011.

[7] W. Min and L. Wynter, "Real-time road traffic prediction with spatio-temporal correlations," *Transportation Research Part C: Emerging Technologies*, vol. 19, pp. 606–616, 2011.

[8] K. Y. Chan, T. Dillon, E. Chang, and J. Singh, "Prediction of Short-Term Traffic Variables Using Intelligent Swarm-Based Neural Networks," *IEEE Transactions on Control Systems Technology*, vol. 21, pp. 263–274, 2013.

[9] B. Ghosh, B. Basu, and M. O'Mahony, "Multivariate Short-Term Traffic Flow Forecasting Using Time-Series Analysis," *IEEE Transactions on Intelligent Transportation Systems*, vol. 10, pp. 246–254, 2009.

[10] Y. Kamarianakis, H. O. Gao, and P. Prastacos, "Characterizing regimes in daily cycles of urban traffic using smooth-transition regressions," *Transportation Research Part C: Emerging Technologies*, vol. 18, pp. 821–840, 2010.

[11] C. Chen, Y. Wang, L. Li, J. Hu, and Z. Zhang, "The retrieval of intra-day trend and its influence on traffic prediction," *Transportation Research Part C: Emerging Technologies*, vol. 22, pp. 103–118, 2012.

[12] H. Dia, "An object-oriented neural network approach to short-term traffic forecasting," *European Journal of Operational Research*, vol. 131, pp. 253–261, 2001.

[13] S. Innamaa, "Short-term prediction of travel time using neural networks on an interurban highway," *Transportation*, vol. 32, pp. 649–669, 2005.

[14] W. H. Lam, Y. Tang, and M.-L. Tam, "Comparison of two non-parametric models for daily traffic forecasting in Hong Kong," *Journal of Forecasting*, vol. 25, pp. 173–192, 2006.

[15] Z. Zheng and D. Su, "Short-term traffic volume forecasting: A k-nearest neighbor approach enhanced by constrained linearly sewing principle component algorithm," *Transportation Research Part C: Emerging Technologies*, vol. 43, pp. 143–157, 2014.

[16] B. Williams, P. Durvasula, and D. Brown, "Urban Freeway Traffic Flow Prediction: Application of Seasonal Autoregressive Integrated Moving Average and Exponential Smoothing Models," *Transportation Research Record*, vol. 1644, pp. 132–141, 1998.

[17] S. Shekhar and B. Williams, "Adaptive Seasonal Time Series Models for Forecasting Short-Term Traffic Flow," *Transportation Research Record*, vol. 2024, pp. 116–125, 2008.

[18] N. Zhang, Y. Zhang, and H. Lu, "Seasonal Autore-

gressive Integrated Moving Average and Support Vector Machine Models," *Transportation Research Record*, vol. 2215, pp. 85–92, 2011.
[19] H. Tan, Q. Li, Y. Wu, W. Wang, and B. Ran, "Freeway Short-Term Travel Time Prediction Based on Dynamic Tensor Completion," *Transportation Research Record*, vol. 2489, pp. 97–104, 2015.
[20] Y. Lv, Y. Duan, W. Kang, Z. Li, and F.-Y. Wang, "Traffic Flow Prediction with Big Data: A Deep Learning Approach," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, pp. 865–873, 2015.
[21] N. G. Polson and V. O. Sokolov, "Deep learning for short-term traffic flow prediction," *Transportation Research Part C: Emerging Technologies*, vol. 79, pp. 1–17, 2017.
[22] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *The Annals of Statistics*, vol. 29, pp. 1189–1232, 10 2001.
[23] R. J. Hyndman, R. A. Ahmed, G. Athanasopoulos, and H. L. Shang, "Optimal combination forecasts for hierarchical time series," *Computational Statistics & Data Analysis*, vol. 55, pp. 2579–2589, 2011.
[24] I. Okutani and Y. J. Stephanedes, "Dynamic prediction of traffic volume through Kalman filtering theory," *Transportation Research Part B: Methodological*, vol. 18, pp. 1–11, 1984.
[25] R. Chrobok, O. Kaumann, J. Wahle, and M. Schreckenberg, "Different methods of traffic forecast based on real data," *European Journal of Operational Research*, vol. 155, pp. 558–568, 2004.
[26] S. Kim, "ppcor: an R package for a fast calculation to semi-partial correlation coefficients," *Communications for statistical applications and methods*, vol. 22, p. 665, 2015.
[27] S. Sun, R. Huang, and Y. Gao, "Network-scale traffic modeling and forecasting with graphical lasso and neural networks," *Journal of Transportation Engineering*, vol. 138, pp. 1358–1367, 2012.
[28] J. Z. Zhu, J. X. Cao, and Y. Zhu, "Traffic volume forecasting based on radial basis function neural network with the consideration of traffic flows at the adjacent intersections," *Transportation Research Part C: Emerging Technologies*, vol. 47, pp. 139–154, 2014.
[29] C. Oh, S. Ritchie, and J.-S. Oh, "Exploring the relationship between data aggregation and predictability to provide better predictive traffic information," *Transportation Research Record: Journal of the Transportation Research Board*, pp. 28–36, 2005.
[30] J. Guo, B. Williams, and B. Smith, "Data Collection Time Intervals for Stochastic Short-Term Traffic Flow Forecasting," *Transportation Research Record*, vol. 2024, pp. 18–26, 2008.
[31] E. Vlahogianni and M. Karlaftis, "Temporal aggregation in traffic data: implications for statistical characteristics and model choice," *Transportation Letters*, vol. 3, pp. 37–49, 2011.
[32] P. Bühlmann and B. Yu, "Boosting With the L2 Loss," *Journal of the American Statistical Association*, vol. 98, pp. 324–339, 2003.
[33] P. Bühlmann, "Boosting for high-dimensional linear models," *The Annals of Statistics*, vol. 34, pp. 559–583, 2006.
[34] R. E. Schapire and Y. Freund, *Boosting: Foundations and algorithms*. MIT Press, 2012.
[35] T. J. Hastie and R. J. Tibshirani, *Generalized additive models*. CRC Press, 1990.
[36] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*. Springer series in statistics New York, 2001, vol. 1, no. 10.
[37] A. Clements, A. Hurn, and Z. Li, "Forecasting day-ahead electricity load using a multiple equation time series approach," *European Journal of Operational Research*, vol. 251, pp. 522–530, 2016.
[38] Z. Li, A. Hurn, and A. Clements, "Forecasting quantiles of day-ahead electricity load," *Energy Economics*, vol. 67, pp. 60–71, 2017.
[39] B. M. Williams and L. A. Hoel, "Modeling and Forecasting Vehicular Traffic Flow as a Seasonal ARIMA Process: Theoretical Basis and Empirical Results," *Journal of Transportation Engineering*, vol. 129, pp. 664–672, 2003.
[40] A. Stathopoulos and M. G. Karlaftis, "A multivariate state space approach for urban traffic flow modeling and prediction," *Transportation Research Part C: Emerging Technologies*, vol. 11, pp. 121–135, 2003.
[41] M. Treiber and A. Kesting, "Trajectory and Floating-Car Data," in *Traffic Flow Dynamics*. Springer, Berlin, Heidelberg, 2013, pp. 7–12.
[42] B. L. Smith, B. M. Williams, and R. K. Oswald, "Comparison of parametric and nonparametric models for traffic flow forecasting," *Transportation Research Part C: Emerging Technologies*, vol. 10, pp. 303–321, 2002.
[43] M. G. Karlaftis and E. I. Vlahogianni, "Memory properties and fractional integration in transportation time-series," *Transportation Research Part C: Emerging Technologies*, vol. 17, pp. 444–453, 2009.

APPENDIX. THE SPECIFICATIONS OF THE SARIMA AND KALMAN FILTER MODELS USED IN THE MODEL COMPARISON

SARIMA: Letting the target variable and the seasonally differenced target variable at time t of day d and location l be denoted as $y_{t,d,l}$ and $z_{t,d,l}$, respectively, then the model specification can be written as:

$$\phi(B)\Phi(B^S)z_{t,d,l} = \theta(B)\Theta(B^S)\varepsilon_{h,d,l}.$$

In which, B is the backshift operator and S the seasonality in terms of the number of time periods. The error term, $\varepsilon_{t,d,l} \sim N(0,\sigma^2)$, where $N(0,\sigma^2)$ denotes the normal distribution with mean 0 and variance σ^2.

In the specification used in the comparison of Section V, SARIMA(1,0,1)(0,1,1)$_S$, the equation above translates to:

$$(1-B)z_{t,d,l} = (1-B)(1-B^S)\varepsilon_{h,d,l}.$$

Kalman filter: The target variable $y_{t,d,l}$ can be expressed as a linear combinations of regression parameters, predictors and the error term:

$$y_{t,d,l} = \boldsymbol{X}'_{t,d,l}\boldsymbol{\alpha}_t + \varepsilon_{t,d,l}, \quad \varepsilon_{t,d,l} \sim N(o, \sigma^2).$$

In which, $\boldsymbol{X}_{t,d,l}$ is a column vector containing all the predictors at time t of day d and location l. The predictors contained in $\boldsymbol{X}_{t,d,l}$ for the three datasets are:

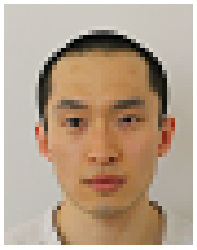
$$\{y_{t-j,d,l} : j = 1, \ldots, 8\} \cup \{y_{t,d-1,l}, y_{t,d-7,l}\} \cup \{\mathbf{1}_{d,l,w} y_{t,d-1,l} : w = 1 \ldots 6\} \cup \{\varepsilon_{t-i,d,l} : i = 1 \ldots 3\}, \quad \text{(Dataset I)}$$

$$\{y_{t-j,d,l} : j = 1, \ldots, 9\} \cup \{y_{t,d-1,l}\} \cup \{\varepsilon_{t-i,d,l} : i = 1 \ldots 3\}, \quad \text{(Dataset II)}$$

$$\{y_{t-j,d,l} : j = 1, \ldots, 20\} \cup \{\varepsilon_{t-i,d,l} : i = 1 \ldots 3\}. \quad \text{(Dataset III)}$$

Letting C be the total number of predictors included in $\boldsymbol{X}_{t,d,l}$, then the parameter vector $\boldsymbol{\alpha}_t = [\alpha_{t,1}, \cdots, \alpha_{t,C}]'$. And for $c = 1, \cdots, C$:

$$\alpha_{t,c} = \phi_c \alpha_{t-1,c} + \epsilon_{t,c}, \quad \epsilon_{t,c} \sim N(0, \sigma_c^2).$$



Zili Li received his PhD degree in economics from Queensland University of Technology, Brisbane, Australia, in 2016. He is currently working as an Post-doctoral research fellow in the School of Civil Engineering, the University of Queensland, Brisbane, Australia.

His research interests include forecasting, operational research and statistical modelling.

Zuduo Zheng is an Associate Professor in the School of Civil Engineering, the University of Queensland, and a DECRA Research Fellow sponsored by the Australian Research Council. His research interests lie primarily in the areas of traffic flow modelling, travel behaviour and decision making, advanced data analysis techniques (e.g., mathematical modelling, econometrics, numerical optimisation) in transport engineering, and meta-research. He is on the editorial advisory boards of several transport journals including Transportation Research Part B, Transportation Research Part C, Heliyon, and International Journal of Intelligent Transportation Systems Research.

Simon Washington is Head of the School of Civil Engineering at the University of Queensland. His research interests include the areas of behavioural econometrics applied to transport safety, travel behaviour, and transport planning. He is on the editorial advisory boards of several transport journals including the Journal of Sustainable Transport, Transportation Research Part A, Journal of Accident Modelling and Research, Korean Journal of Transportation Engineering, Journal of Transport and Health, and Accident Analysis and Prevention.