Space-time modeling of traffic variables with adaptive LASSO

Yiannis Kamarianakis
School of Mathematical and Statistical Sciences, Arizona State University.

Abstract

This article presents a parametric time-series model for short-term traffic forecasting, which accounts for space-time dependencies and cross-correlations of traffic variables. In the adopted framework, a separate model is built for each measurement location in the network. Each model may contain some hundreds of potentially useful predictors, that contain information from other locations in the network; the influential ones are chosen via a 2-step, penalized estimation scheme, namely, adaptive LASSO. In the Athens data, the method achieves approximately 90% 1-step forecast accuracy on 3-minute volumes, which declines slowly as the forecast-horizon increases; the observed forecasting performance for occupancies is not as satisfactory though. An advantage of the proposed approach is that it is readily able to be automated.

Keywords: traffic forecasting; real-time predictions; threshold regressions; adaptive LASSO.

1 Introduction

Accurate short-term forecasting of traffic variables is essential for intelligent transportation systems applications, such as real-time route guidance and advanced traveler information systems. Hence, numerous modeling approaches have been proposed during the last two decades [1, 2], including both nonparametric and parametric models, the latter being nonlinear in the vast majority of recent studies. Most published research works focus on point-wise forecast accuracy; construction of confidence bands and issues such as computability and scalability of a proposed modeling framework are frequently omitted from the discussion. Typically, short-term traffic forecasting models are evaluated on data from a couple of dozens of measurement locations at most, the evaluation periods being usually a few days. Incident detection and forecasting performance in the presence of incidents have not received much attention, despite of being critical in real-world implementation of real-time traffic forecasting systems.

Traffic forecasting models are usually evaluated on data from arterials and freeways, which are admittedly less variable than data from urban networks and not subject to the effects of traffic lights. In urban networks, neighborhood relationships and the definitions of spatial weight matrices for space-time parametric frameworks as the ones presented in [3, 4], are not straightforward; some locations may not be clearly upstream or downstream a given location. Furthermore, detectors can be dense in an urban network, so that locations with useful predictive information may be hard to identify; this again affects the construction of spatial weight matrices used in space-time modeling schemes. Erroneous and missing data are expected to be more frequent in urban networks, which makes essential the implementation of robust estimation procedures.

This article presents an approach to short-term vehicular traffic forecasting that is based on parametric spatial time-series models, and discusses its performance on the data provided from the organizers of the Transportation Data Forecasting Competition. The applied models are variants of the ones presented in [5] which take into account cross-corelations of traffic variables in the spirit of [6]. It is worth noting that [5] reported accuracies (accuracy defined as 1-mean absolute percentage error) greater than 90% in highways and urban arterials. The adopted modeling framework avoids issues which can be challenging in real-world traffic forecasting applications, that involve numerous measurement locations. First, it does not require a model building procedure that uses computationally expensive sequential statistical tests [7, 8] to identify the number of traffic regimes. Second, nonlinear parametric estimation, which can be sensitive to the choice of initial values and computationally intensive, is avoided as well. Third, the modeling framework does not demand the a-priori specification of spatial weight matrices (as for instance, in [3, 4]) for the representation of neighboring associations among measurement locations; as shown for example in [9], construction of such matrices and statistical evaluation of their alternative forms can be a tedious step in the model-building procedure.

The above mentioned gains come at the cost of a substantially increased number of predictors in the linear specification. The influential ones are identified by a two-step penalized estimation scheme, namely adaptive least absolute shrinkage and selection operator (LASSO); for recent applications of penalized estimation in transportation problems, the reader may consult [10–12]. In the forecasting experiments we combine models estimated using: (i) the adaptive LASSO which performs $l1$-penalized minimization of squared residuals and (ii) the adaptive LAD-LASSO which produces $l1$-penalized least absolute deviation estimators. The latter are essentially median regression estimates which have been found to be particularly effective in terms of forecasting performance when response variables possess skewed response distributions that may contain outliers [13].

The next section presents the methodology. Section 3 discusses details on model building and forecasting performance on the training data. Finally, we present some concluding remarks and recommendations for further research.

2 Modeling traffic variables with space-time regressions and penalized estimation

2.1 Decomposition of traffic dynamics

Our modeling strategy is based on a variant of the unobserved components time-series model class. The decomposition of an observed time series into unobserved stochastic processes may provide improved understanding of the dynamic characteristics of the series and their temporal evolution [14]. Specifically, let i be the traffic-variable index, s the location index and t the time of day index. The overall day-specific model structure for traffic variable y_i is

$$y_{i,s}(t) = \mu_{i,s}(t) + x_{i,s}(t) \tag{1}$$

where $i = 1,...,I$, $s = 1,...,S$ and $t = 1...,T$. I represents the number of traffic variables, S the number of measurement locations and T is the number of time intervals per day, which depends on the level of temporal aggregation. The $\mu_{i,s}$ capture the location-specific daily profiles and can be viewed as baseline predictors that use only historical data and neglect information from the recent past of the process. Each day of the week has its own profile in (1); hence at this stage of the analysis we do not perform clustering of days with similar dynamics. A method based on curve-clustering and functional data analysis has been proposed in [15]. A variety of methods have been proposed for the estimation of $\mu_{i,s}(t)$; in the application we adopted the method proposed in [16].

2.2 The transient model

The second term in the right hand side of (1) represents the dynamics of the short-term deviations from the historical daily profiles; it is modeled by the most parsimonious nonlinear specification, namely threshold regression. To illustrate the model, we assume that a single variable governs the transition between different (linear) regimes; the modeling framework can be extended to multiple such variables in a straightforward manner. In particular, for each traffic variable i, and for each location s, a space-time threshold autoregressive model is employed to capture transient behavior:

$$x_{i,d,s}(t) = \alpha_{i,d,s}^{\left(r_{i,d,s}\right)} + \sum_{i=1}^{I}\sum_{k=1}^{p} \alpha_{i,k,d,s}^{\left(r_{i,d,s}\right)} x_{i,d,s}(t-k) + \sum_{i=1}^{I}\sum_{j=1}^{N_s}\sum_{k=1}^{p} \alpha_{i,j,k,d,s}^{\left(r_{i,d,s}\right)} x_{i,j,k,d,s}(t-k) + \varepsilon_{i,d,s}(t). \quad (2)$$

In (2), $d = 1,...,D$ is an index that specifies groups of days; D may be less than seven if there is sufficient evidence of similarity of traffic dynamics for two (or more) days of the week. The operating regime in the piecewise linear formulation is specified by $r_{i,d,s} = 1,...,R_{i,d,s}+1$; regimes are characterized by an increasing sequence of thresholds and the values of the chosen transition variable. Hence, (2) defines a threshold regression per measurement location, with an unknown number of regimes, as in [7]. In the application, time-of-day is the threshold variable that defines subsamples in which the relationship is stable. In general the threshold variable can be subject to a model building procedure which chooses the traffic variable for which linearity is more strongly rejected; such a procedure is presented in [17].

The predictive model contains an intercept term that varies with the traffic variable modeled, with location, day-of-the-week and traffic regime within a day. The two sums in (2) contain information on past deviations of traffic variables from their historical profiles. N_s is the number of neighboring locations of s that may contain useful past information with regard to short-term forecasting performance and p is the autoregressive order (maximum time lag) of the model. The alphas are unknown coefficients; the ones which are significantly different from zero signify which temporal lags of each neighboring location provide useful information with regard to short-term forecasting. The last term in (2), is assumed to be a martingale difference sequence with respect to the history of the time-series up to time $t-1$; hence, it is assumed a serially uncorrelated (but not necessarily independent) sequence and its variance is not restricted to be equal across different regimes.

Despite having a dataset with a small spatial dimension in the application, we do not consider simultaneous estimation of a system of equations with a common covariance matrix (each equation corresponding to a measurement location in the network) as for instance in [3, 4]. Although a system in general is expected to produce more efficient estimates compared to the equation-by-equation approach, such an estimation framework cannot be applied in practice when S is as large as in real-world applications, where S may easily exceed 300.

2.3 Penalized estimation for automatic model selection

Direct estimation of models as the one presented above using conventional least squares or least absolute deviation algorithms, is expected to be inefficient as a fraction of the predictors will not contribute significantly to the predictive power of the model. In some cases direct estimation may be problematic, with unacceptably high variances for the estimated coefficients, or even infeasible because of multicollinearity. To avoid these problems, a penalized estimation scheme that performs simultaneous estimation and predictor selection for each location is used. Penalized estimation is preferred to stepwise model building based on statistical information criteria (e.g. AIC or BIC) which can be overly demanding in terms of computational power for large S.

Specifically, given a loss function $g(.)$, coefficient estimation within regime $r_{i,d,s}$ in (2) is performed by minimizing the criterion:

$$f(\varepsilon) = g(\varepsilon) + \sum_{i=1}^{I}\sum_{k=1}^{p} \lambda_{i,k,d,s}^{(r_{i,d,s})} \left| \alpha_{i,k,d,s}^{(r_{i,d,s})} \right| + \sum_{i=1}^{I}\sum_{j=1}^{N_s}\sum_{k=1}^{p} \lambda_{i,j,k,d,s}^{(r_{i,d,s})} \left| \alpha_{i,j,k,d,s}^{(r_{i,d,s})} \right|. \quad (3)$$

In the application two criteria were used: the first, henceforth referred to as LAD-LASSO, minimizes the sum of absolute residuals in the right hand side of (3); the second minimizes a least-squares objective and is referred to as LASSO. Hence, in the former case g is the absolute value function whereas in the latter, g is quadratic. The lambdas in (3) are penalty terms which shrink coefficients toward the origin and tend to discourage models with large numbers of marginally relevant predictors. In the application the penalty terms, which are equal in number to the unknown coefficients, are selected as inversely proportional to initial, first-stage coefficient estimates derived using slight penalization, following the procedure suggested in [18, 19].

One expects that, depending on the characteristics of traffic data and the density of measurement locations in a road network, there is a maximum time lag and a maximum number of neighbors, above which additional predictors in a linear model do not contribute in terms of out-of-sample predictive ability. When p and N_s take very large values the estimation problem becomes harder to solve and the finite sample performance of the estimator degrades slightly; consequently the out-of-sample predictive ability weakens. It is worth noting that by using different versions of the check function in defining g in (3), one obtains penalized quantile regression estimates in a straightforward manner [20]; such estimates can be used for the construction of confidence bands around the point forecasts.

3 The Application

3.1 Model calibration

The proposed modeling strategy is applied to traffic volumes and occupancies collected from seven measurement locations in Alexandras Ave., Athens, at two levels of temporal aggregation: 3- and 15-minute. It is worth emphasizing that in most applications, multiperiod forecasting involves the use of a chain rule to generate forecasts at longer horizons, based on a dynamic model for data observed at a high-frequency level (e.g., 1.5 minutes in this case). Under this iterated approach model, specification is the same across all forecast horizons; only the number of iterations changes with the horizon. The approach that was chosen here on the other hand, allows for varying forecasting models across different horizons. As noted in [21] both approaches have advantages and drawbacks: "For a given model specification the iterated approach leads to more efficient parameter estimates since it includes data recorded at the highest available frequency and so uses the largest available sample size. If the model is misspecified, due, for example, to an omitted variable or because of an incorrect lag order, iterating the model multiple steps ahead can either attenuate or reduce existing biases. Direct forecasts are less efficient, but also more likely to be robust to model misspecification as they are typically linear projections of current realizations on past data."

Calibration of the parameter settings in (2), was based on a cross-validation experiment which used 2 testing weeks (starting on May 7 and May 14, 2000). Alternative parameterizations were evaluated using 4 training weeks and the chosen performance metrics were: mean absolute error (MAE) for occupancies and mean absolute percentage error (MAPE) for volumes. To keep the analysis simple, parameterizations were allowed to vary per traffic variable, but not per measurement location or day-of-the-week. Specifically, for models estimated on 3-minute data the evaluated parameterizations were:

— The maximum autoregressive order p which ranged from 2 to 14.

— The day groups: $D = 1$ when the transient models are equal across days, $D = 2$ for different models in weekends (versus weekdays), $D = 3$ when weekdays are separated into two groups with Tuesday, Wednesday and Thursday in one group and $D = 7$ for day-specific transient models.

— The temporal regimes: $R_i = 0$ corresponds to a single regime, $R_i = 1$ was tested as a day-(starting at 05:30 AM) versus night-regime (starting at 21:30), when $R_i = 2$ the day-regime was further divided into a morning-peak period that lasted till noon and an afternoon-peak, whereas when $R_i = 3$ an intermediate regime that lasted from 10:30 AM till 15:00 PM, divided the two peak-periods.

— the maximum number of neighbors: $N_s = 0$ when no neighboring information is included in the transient models, $N_s = 3$ when neighbors are locations with the same direction as the location modeled, and $N_s = 6$ when all locations are neighbors.

— The static weight ρ, for LAD-LASSO (versus LASSO) forecasts, which ranged from 0 to 1, in steps of 0.2.

The evaluated parameterizations for the 15-minute data were different with regard to: a) the autoregressive orders which ranged from 1 to 6 and b) the temporal regimes (multi-regime models were not tested in this case). It should be stressed that some of the model settings depicted above (for instance the definition of temporal regimes) were based on prior experience from other traffic-forecasting applications. Prior knowledge also suggested that forecasting performance improves with the length of the training period (at least for training periods that are not longer than 12 weeks); hence, the whole amount of training data was used in the calibration of the final models.

3.2 Model evaluation

Optimal forecasting performance for models estimated using 3-minute volumes was achieved with $p = 14$, $D = 1$, $R_i = 1$, $N_s = 3$ and $\rho = 0.6$. It should be noted that this choice was based on evaluations of accuracies for multiple horizons; for instance, $p = 10$ optimizes 1-step performance and $p = 14$ is optimal for multiple-step forecasts. Since the choice $p = 14$ had a negligible effect on 1-step forecasts it was decided to estimate the final 3-minute models using the latter parameterization. Although multiple-regime specifications could be expected to be superior a-priori, they result into reduced sample sizes which may have a negative effect on the performance of the adaptive LASSO algorithm. Median 1-step accuracy across measurement locations (defined as 1-MAPE) approached 90% and decreased slowly as the horizon increased, with 87% median accuracy for 15-minute forecasts. After 5 forecasting steps, the transient models are not useful as they cannot beat the baseline (which is based solely on the estimated profiles). The worst 1-step performance (84.7%)was observed for location L-104 whereas the best (91.3%) for L-101.

On the other hand, forecasting performance for models estimated on 3-minute occupancies was achieved with $p = 8$, $D = 1$, $R_i = 1$, $N_s = 3$ and $\rho = 0.8$. Median 1-step MAE across measurement locations was close to 8 and increased relatively fast as the horizon increased, with median MAE= 12.2 for 15-minute forecasts. After 8 forecasting steps, the transient models are not useful as they cannot beat the baseline (which achieves median MAE= 14.4). The worst 1-step performance (MAE= 11.46)was observed for location L-103 whereas the best (MAE= 5.3) for L-107.

Optimal performance for models estimated using 15-minute volumes and occupancies was achieved with $p = 4$, $D = 1$, $N_s = 3$ and $\rho = 0.6$. 15-minute aggregation smooths traffic data; in this case, median 1-step (15-minute ahead forecast) accuracies for volumes approached 92%, median 4-step ahead forecasts approached 90% whereas median 1-step MAE for occupancies was close to 9 and median 4-step MAE approached 13.2.

Although evaluation of models that use data with alternative temporal aggregation intervals should have been perfromed on the same basis, (for instance using the observed data at 1.5 minute frequency) this step was not performed due to temporal limitations. It was decided that 1.5-minute ahead forecasts at time t would be created as averages of the current data and the 3-minute forecasts; similarly, 15-minute forecasts were produced by averaging 5-step forecasts from the 3-minute models with one-step forecasts from the 15-minute models. Finally, forecasts for longer horizons were created by models calibrated solely on 15-minute aggregated data. It should be stressed that an alternative approach (that is expected to be more accurate) could be to employ a less ‘naive’ forecast combination scheme that could perhaps extend to horizons larger than 15-minute, as in [22].

4 Concluding Remarks

This work has several limitations. One of them was that an incident detection algorithm was not applied: incident-related and erroneous data should have been removed before calibrating our models. In the presence of incidents, it is expected that optimal forecasting models are significantly different.

Furthermore, in challenging real-time traffic forecasting problems such as those that arise in forecasting urban traffic, the question as to which is the best strategy may be effectively expanded and answered via the use of a (possibly regime-switching) forecast combination scheme. Ideally the solution that is presented here would have been combined with alternative modeling approaches, such as Support Vector Regressions.

Despite the above-mentioned limitations, an advantage of the proposed solution is that it is readily able to be automated for large-scale problems. Performance similar to the one observed here, was also observed in other urban networks with several hundreds measurement locations.

Acknowledgments

The author wishes to thank Laura Wynter for helpful discussions and to congratulate the organizers of the TRANSFOR competition for their initiative.

References

[1] Vlahogianni EI, Golias JC, Karlaftis MG. Short-term Traffic Forecasting: Overview of Objectives and Methods. *Transport Reviews* 2006, **24**: 533-558.

[2] Karlaftis MG, Vlahogianni EI. Statistical methods versus neural networks in transportation research: Differences, similarities and some insights. *Transportation Research Part C* 2011, **19**: 387-399.

[3] Kamarianakis Y, Prastacos P. Space-time modeling of traffic flow. *Computers* & *Geosciences* 2005, **31**: 119-133.

[4] Min W, Wynter L. Real-time road traffic prediction with spatio-temporal correlations. *Transportation Research Part C* 2011, **19**: 606-616.

[5] Kamarianakis Y, Shen W, Wynter L. Real-time road traffic forecasting using regime-switching space-time models and adaptive LASSO. *Applied Stochastic Models in Business and Industry* 2012, **28**: 297-315.

[6] Tavana H, Mahmassani S. Estimation and application of dynamic speed-density relations by using transfer function models. *Transportation Research Record-Journal of the Transportation Research Board* 2000, **1710**: 47-57.

[7] Hansen BE. Sample splitting and threshold estimation. *Econometrica* 2000, **68**: 575-603.

[8] Strikholm B, Tersvirta T. A sequential procedure for determining the number of regimes in a threshold autoregressive model. *Econometrics Journal* 2006, **9**: 472-491.

[9] Anselin, L. Non-nested tests on the weight structure in spatial autoregressive models: some monte-carlo results. *Journal of Regional Science* 1986, **26**: 267-284.

[10] Gao Y, Sun S, Shi D. Network-scale traffic modeling and forecasting with graphical lasso. *Advances in Neural Networks-ISSN* 2011, Lecture Notes in Computer Science Vol. 6676, 151-158.

[11] Hofleitner A, El Ghaoui L, Bayen A. Online least-squares estimation of time varying systems with sparse temporal evolution and application to traffic estimation. *50th IEEE Conference on Decision and Control* 2011, 2595-2601.

[12] Wan K, Kornhauser A. Link-data-based approximation of path travel time distribution with Gaussian copula estimated through LASSO. *Transportation Research Board Annual Meeting* 2010, Paper 10-2769.

[13] Dielman T. A comparison of forecasts from least absolute value and least squares regression. *Journal of Forecasting* 1986, **5**: 189-195.

[14] Koopman SJ, Ooms M. Forecasting economic time series using the unobserved components time series models. *The Oxford Handbook of Economic Forecasting* 2011, Clements and Hendry (eds), Oxford University Press, 124-162.

[15] Garcia-Escudero LA, Gondaliza A. A proposal for robust curve clustering. *Journal of Classification* 2005, **22**: 185-201.

[16] Cleveland RB, Cleveland WS, McRae JE, Terpenning I. STL: a seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics* 1990, **6**: 3-73.

[17] Kamarianakis Y, Gao HO, Prastacos P. Characterizing regimes in daily cycles of urban traffic using smooth-transition regressions. *Transportation Research Part C* 2010, **18**: 821-840.

[18] Wang H, Li G, Jiang G. Robust regression shrinkage and consistent variable selection through the LAD-LASSO. *Journal of Business and Economic Statistics* 2007, **25**: 347-355.

[19] Pötscher BM, Schneider U. On the distribution of the adaptive LASSO estimator. *Journal of Statistical Planning and Inference* 2009, **139**: 2775-2790.

[20] Koenker R. *Quantile Regression*; Cambridge University Press, 2005.

[21] Pesaran MH, Pick A, Timmermann AG. Variable selection and inference for multi-period forecasting problems. *Journal of Econometrics* 2011, **167**: 173-187.

[22] Andrawis RR, Atiya AF, El-Shishiny H. Combination of long term and short term forecasts, with application to tourism demand forecasting. *International Journal of Forecasting* 2011, **27**: 870-886.