ROUTE TRAVEL TIME PREDICTION AND VARIABILITY ESTIMATION IN AN URBAN NETWORK BASED ON TAXI FLOATING CAR DATA

MIRSAD TULIC

Austrian Institute of Technology, Mobility Department, Dynamic Transportation Systems

DIETMAR BAUER

Austrian Institute of Technology, Mobility Department, Dynamic Transportation Systems

WOLFGANG SCHERRER

Vienna University of Technology, Institute for Mathematical Methods in Economics, Research Unit Econometrics and Systemtheory

1. INTRODUCTION

Real time information systems providing estimates of expected route travel times are already commonplace. These services typically do not provide estimates of travel time variability, although this is important for applications with different costs for coming late or early.

Clearly the prediction of route travel time simply is the sum of the predicted link travel times. For the variability of the prediction, however, the full variance-covariance matrix needs to be considered. Moreover, link travel times show temporal and spatial dependence as well as heteroskedasticity in the sense that link travel time variability depends on traffic conditions.

Such features have been considered already in the literature: E.g. in [1] and [2] the spatial dependence of dynamic traffic variables are incorporated, but heteroskedasticity of the one step prediction errors is neglected. [2] incorporates spatial dependence for link travel time predictions based solely on geographical closeness.

In contrast to earlier literature in this paper an automatic method to obtain short-term predictions of route travel time and the corresponding variability is presented taking spatial and temporal correlations as well as observed heteroskedasticity of link travel times into account. The method suggested below make extensive use of model selection techniques and hence can be applied to a large number of links making routing applications in urban regions possible.

2. DATA DESCRIPTION

This paper uses the taxi floating car data in Vienna (see [3] for details) comprising a fleet of approximately 3500 taxis with a sampling frequency between 30 to 60 seconds. The data comprises harmonic averages of link speeds $v_{l,i}(t)$ on link l in time interval i (15-minute intervals) on day t. A selection of 75 links from July 1st 2008 to July 31st 2010 in the vicinity of the Museumsquartier (see Figure 1) are used. About 39% of the data is missing. The data set was split into training data (July 1st 2008 to June 2nd 2010) and validation data (the remaining 60 days).

E-mail addresses: Mirsad.Tulic.fl@ait.ac.at.

3. THE MODEL

The variable $y_{l,i}(t)$ (link speed $v_{l,i}(t)$ or link travel time $TT_{l,i}(t)$) for link l, a time-of-the-day 15-minute interval i and day t is modelled as a random variable:

$$y_{l,i}(t) = \mu_{l,i}(t) + \varepsilon_{l,i}(t) = \mathbf{X}(t)'\beta_{l,i} + \varepsilon_{l,i}(t), \;\; t = 1,\ldots,T, i = 1,\ldots,J, l = 1,\ldots,L, \tag{1}$$

where $\mathbf{X}(t) \in \mathbb{R}^k, \beta_{l,i} \in \mathbb{R}^k$, $\mu_{l,i}(t) = \mathbb{E}\left[y_{l,i}(t)\right]$ where $\beta_{l,i}$ does not depend on the day. $\mathbf{X}(t)$ models effects of the weekday and seasonal effects and thus contains Fourier frequencies $\cos(k \cdot \frac{2\pi \cdot t}{365}), \sin(k \cdot \frac{2\pi \cdot t}{365})$ weekday dummy variables and a school-holiday dummy. Note that the daily profile is not represented in the regressors but in $\beta_{l,i}$ depending on the time-of-day interval.

To account for heteroscedasticity due to traffic conditions and aggregation of the speeds of the observed taxis to obtain average speeds for each time interval, the variance of the residuals $\varepsilon_{l,i}(t)$ is modelled as a function of the number, $n_{l,i}(t)$ say, of taxis contributing, and the current average link speed or travel time respectively:

$$\mathbb{E}\varepsilon_{l,i}^2(t) = \sigma_{l,i}^2(t) = \exp\left(\alpha_{l,i} + \gamma_{l,i}\frac{1}{\sqrt{n_{l,i}(t)}} + \delta_{l,i}\mu_{l,i}(t)\right). \tag{2}$$

In order to model the spatial and temporal correlations the following autoregressive specification is used:

$$\varepsilon_{l,i}(t) = \Gamma_l \varepsilon_{:,i-1}(t) + \zeta_l(t), \;\; t = 1,\ldots,T, i = 1,\ldots,J, l = 1,\ldots,L, \tag{3}$$

where $\Gamma_l \in \mathbb{R}^L$ and $\varepsilon_{:,i-1}(t)$ equals the vector obtained by stacking the residuals for all links where $\varepsilon_{:,0}(t) := \varepsilon_{:,96}(t-1)$.

The one-step ahead predictions

$$\hat{y}_{l,i+1}(t) = \mathbf{X}(t)\beta_l + \Gamma_l \cdot \varepsilon_{:,i}(t). \tag{4}$$

Alternatively in place of using an autoregressive model for the residuals $\varepsilon_{:,i}(t)$ the model can be formulated using the normalized residuals $\varepsilon_{l,i}^*(t) := \sigma_{l,i}^{-1}(t)\varepsilon_{l,i}(t)$:

$$\varepsilon_{l,i}^*(t) = \Gamma_l^* \varepsilon_{:,i-1}^*(t) + \zeta_l^*(t), \;\; t = 1,\ldots,T, i = 1,\ldots,J, l = 1,\ldots,L, \tag{5}$$

leading to the one step predictions

$$\hat{y}_{l,i+1}^*(t) = \mathbf{X}(t)\beta_l + \sigma_{l,i+1}(t)\Gamma_l^* \cdot \varepsilon_{:,i}^*(t), \tag{6}$$

Finally the one-step ahead predictions for the link travel times can be summed to obtain route travel time predictions. The variance of the route travel time predictions equals the sum over all entries of the variance-covariance matrix of the link travel time predictions which is modeled as a function of the time-of-the-day.

The choices $y_{l,i}(t) = v_{l,i}(t)$ (and then converting the achieved predicted link speeds to the corresponding link travel times), and $y_{l,i}(t) = TT_{l,i}(t)$ in combination with the two options of unnormalized and normalized residuals provide four alternatives.

4. METHODS FOR AUTOMATIC MODELING

The model described in the previous section potentially contains a large number of parameters due to the fact that not all components of the regressor vector $\mathbf{X}(t)$ are important for each (link, time interval) combination. Also the number of links to include in (3) needs to be specified.

The model selection for $\mu_{l,i}(t)$ is conducted using feasible generalized least squares based on the weights $\hat{\sigma}_{l,i}^{-1}(t)$ and delivers the final set of regressors, estimated coefficients $\tilde{\beta}_{l,i}(t)$, and the residuals $\tilde{\varepsilon}_{l,i}(t) = y_{l,i}(t) - \mathbf{X}(t)'\tilde{\beta}_{l,i}$. A combination of forward selection and backward elimination using information criteria is used here.

The model for the variance $\hat{\sigma}_{l,i}^2(t)$ is based on the preliminary ordinary least squares estimate of the residuals $\hat{\varepsilon}_{l,i}(t) = y_{l,i}(t) - \mathbf{X}(t)'\hat{\beta}_{l,i}$ using the following regression

$$\log\left(\hat{\varepsilon}_{l,i}^2(t)\right) = \alpha_{l,i} + \gamma_{l,i}\frac{1}{\sqrt{n_{l,i}(t)}} + \delta_{l,i}\mathbf{X}(t)'\hat{\beta}_{l,i} + \eta_{l,i}(t).$$

The weights $\hat{\sigma}_{l,i}^2(t)$ are then obtained from (2) by replacing α, γ and δ by their corresponding estimated parameters.

To estimate (3) (or (5)) backward model selection is again applied starting with the set of "neighbours" $\tilde{\varepsilon}_{k,i-1}(t)$ of link l, links lying geographically in a box in the vicinity of link l as well as links being part of often used routes leading through link l. It has been found that including such links not in the vicinity contributes to the prediction accuracy.

type of residuals	$\tilde{\varepsilon}_l(\cdot)$	$\tilde{\varepsilon}^*_l(\cdot)$
mean via travel **speeds** (weighted mean)	6.3292 (8.4738)	6.4567 (8.7202)
mean via travel **times** (weighted mean)	6.0510 (8.1754)	5.9956 (8.0874)

TABLE 1. Results of the performance measures for the prediction of travel times in form of the (weighted) *mean over links* of **RMSE** depending whether the prediction was obtained via travel speeds or directly travel times.

Route No. & Measure	1, RMSE	1, MAPE	2, RMSE	2, MAPE
directly via travel times	11.2794	21.3482	15.5484	21.3984
via travel speeds	10.9687	25.6785	14.5390	23.5730

TABLE 2. Results of the performance measures for the prediction of **route travel times** in form of the RMSE and MAPE depending whether the prediction was obtained via travel speeds or directly travel times. In both cases $\tilde{\varepsilon}^*_l(\cdot)$ are considered. In red colour are the better performing variants (in the sense of the proposed measures)

5. RESULTS

The link-based prediction performance measures $RMSE(l)$ are summarized taking the mean over the 75 links and link-length-weighted means respectively, see Table 1. According to RMSE the method based on travel times is preferred to the method using speeds.

The model obtained directly via travel times, using normalized residuals (TTn) was compared to a simple benchmark model (BM) obtained from regressing $\tilde{\varepsilon}^*_{l,i}(t)$ on $\tilde{\varepsilon}^*_{l,i-1}(t)$. Our model (RMSE: 5.90) narrowly beats the BM (RMSE: 5.91) while achieving superior results in 46 of the 75 links.

Additionally to link-based measures, route travel time performance measures were calculated for two routes (see the green and red lines in Figure 1).

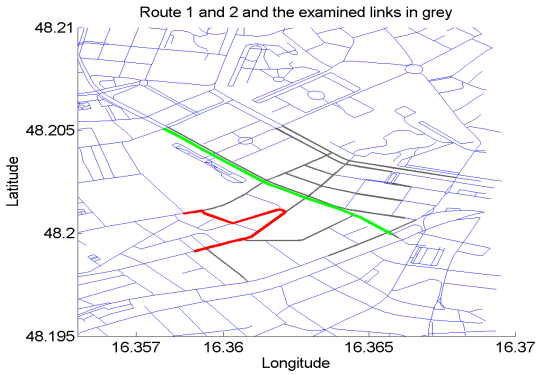

FIGURE 1. The examined area around the street "Getreidemarkt" in Vienna with Route 1 (9 links) in green, Route 2 in red (ten links) and the rest of the 75 links in grey.

The uncertainty of the estimated route travel time prediction for one time interval can be obtained by taking the empirical variance-covariance matrix of the vectors $\mathbf{y}_{k(R),i}$, where $k(R)$ indicates the links constituting the route. An estimate for the uncertainty of the predicted route travel time is thus obtained for each interval i. Figure 2 depicts the described trajectories over the 15-minute intervals of the day for route 1. It is clearly visible that the variability varies over the time-of-the-day with a maximum in the evening peak. Additionally the curves for the estimation and the validation period match indicating the suitability of the suggested model.

REFERENCES

[1] Y. Kamarianakis, W. Shen, and L. Wynter. Y. Real-time road traffic forecasting using regime-switching space-time models and adaptive lasso. to appear in Applied Statistical Journal, 2012.

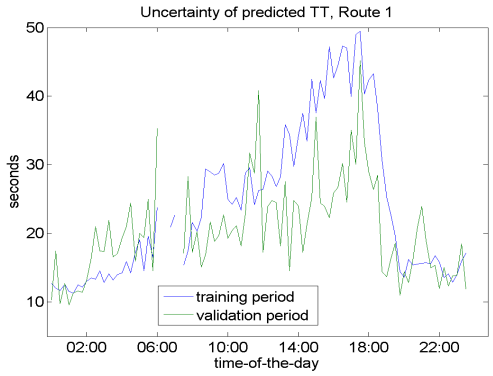

FIGURE 2. The uncertainty of the predicted route travel time for eahc of the 96 quarter-hour intervals of a day. It can be spotted how it rises from the early morning hours to reach a peak in the evening and to drop off from there towards midnight.

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

[3] W. Toplak, H. Koller, M. Dragaschnig, D. Bauer, and J. Asamer. Novel Road Classifications for Large Scale Traffic Networks. In *Proceedings of the ITSC 2010 conference*, Madeira, Portugal, 2010.