

# A meta-model for estimating error bounds in real-time traffic prediction systems

Francisco Pereira, *Member, IEEE*, Constantinos Antoniou, *Member, IEEE*, Joan Aguilar Fargas, and Moshe Ben-Akiva

*AUTHOR'S DRAFT - FOR THE FINAL VERSION, PLEASE GO TO  
<http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=06739140>*

**Abstract**—This paper presents a methodology for estimating the upper and lower bounds of a real-time traffic prediction system, i.e. its prediction interval (PI). Without a very complex implementation work, our model is able to complement any pre-existing prediction system with extra uncertainty information such as the 5% and 95% quantiles. We treat the traffic prediction system as a black box that provides a feed of predictions. Having this feed together with observed values, we then train conditional quantile regression methods that estimate upper and lower quantiles of the error.

The goal of conditional quantile regression is to determine a function,  $d^\tau(\mathbf{x})$ , that returns the specific quantile  $\tau$  of a target variable  $\mathbf{d}$ , given an input vector  $\mathbf{x}$ . Following Koenker [1], we implement two functional forms of  $d^\tau(\mathbf{x})$ : locally weighted linear, which relies on value on the neighborhood of  $\mathbf{x}$ ; and splines, a piecewise defined smooth polynomial function.

We demonstrate this methodology with three different traffic prediction models applied to two freeway data-sets from Irvine, CA, and Tel Aviv in Israel. We contrast the results with a traditional confidence intervals approach that assumes that error is normally distributed with constant (homoscedastic) variance. We apply several evaluation measures based on earlier literature and also contribute two new measures that focus on relative interval length and balance between accuracy and interval length. For the available dataset, we verified that conditional quantile regression outperforms the homoscedastic baseline in the vast majority of the indicators.

Keywords: uncertainty, prediction intervals, dynamic traffic assignment, quantile regression, traffic prediction

## I. INTRODUCTION

Real time traffic prediction involves a very large number of interacting factors, many of them not directly observable, such as noise/no data in sensors and communications, behavioral parameters (e.g. trip destinations), uncertainty in incidents (e.g. actual occurrence times, capacity reduction), lack of detailed weather information and many others. Continuously predicting the state of such a complex system that involves so many spatial and temporal correlations thus becomes prone to a very heterogeneous error structure. As a consequence, traditional error treatment may lead to disappointing estimates and a frustrating service to the end user. Providing confidence intervals that assume a constant (homoscedastic) normal error

distribution may dramatically fail whenever such assumption is violated.

Provision of robust interval estimates is particularly relevant in this context. Traffic prediction systems have the ultimate role of increasing efficiency of the transportation system as a whole, and users' compliance and trust in traffic information and guidance are key to its success. Together with consistent predictions, that take into account driver's reaction to information [2], [3], trust is also supported by accurate uncertainty measures. If the system can't give precise estimates at a given moment, it should not mislead users into believing so. And opportunities to show high accuracy should not be lost either. This can only be achieved with a careful treatment of error.

This paper introduces a methodology for estimating the upper and lower bounds of the error for a real-time traffic prediction system. Since we apply a meta-model perspective, where our prediction model is seen as a black box, this research can be applied to any other prediction model, within and beyond the transportation context.

We define error as the difference between the observation and the prediction. The goal is to estimate the *prediction interval* (PI), i.e. upper and lower bound of this error. Thus, for each prediction that the traffic prediction engine makes by using its own input (e.g. traffic surveillance data, historical data, incident information), our algorithm will make its own prediction of the respective bounds by accessing the same information, potentially together with other sources (e.g. special events, weather, previous error values).

Given the nature of the problem, we want to avoid assumptions about the error form, for example we do not intend to fit it into a gaussian conditioned on the neighborhood, as is sometimes done for heteroscedascity contexts (e.g. [4], [5]). Given the black box assumption, we cannot also introduce variance heterogeneity explicitly into the original model, which is, to our knowledge, the most popular treatment (e.g. [6], [7], [8]). Although an attractive option, propagating individualized error variances back into the traffic prediction model would potentially demand restructuring of the model itself, which is often not an option. Conversely, by using a meta-model, we can develop a relatively complex model that includes context such as incidents, special events or weather.

In order to estimate the bounds, we use conditional quantile regression (see e.g. [1], [9]) which directly estimates a given quantile  $\tau$  of a response variable as a function of an input vector. In our case, the response variable will be the deviation,

F. Pereira is with the Singapore-MIT Alliance for Research and Technology, SMART, Singapore, e-mail: camara@smart.mit.edu.

C. Antoniou is with the National Technical University of Athens, Greece.

J. Fargas and M. Ben-Akiva are with the Massachusetts Institute of Technology (MIT), USA.

$d$ , from the observation, having  $d = y - \hat{y}$ , where  $y$  is the observed value and  $\hat{y}$  is the prediction according to a *black box* prediction model.

Our task is thus to add, to each prediction as provided by a black box model, its upper and lower deviation to obtain a  $Q\%$  prediction interval. In other words, the “observed” values shall fall within such interval at least  $Q\%$  of the times.

For this article, we will focus on prediction of speeds according to three distinct models proposed by Antoniou et al. [10], which incorporate locally weighted scatterplot smoothing (LOESS) regression; multilayer perceptron (Neural Network); and a conventional speed density relationship function, as defined in Ben-Akiva et al. [3]. We will apply such models to speed predictions in two freeway datasets from Irvine, CA, and Tel Aviv in Israel. Each dataset comprises 5 days of data.

Each prediction from these black box models will be augmented with an upper and lower bound estimated by our *meta-model*. For this meta-model, we will test two quantile regression approaches: locally weighted conditional quantile regression, which essentially estimates an independent linear model for each input vector  $\mathbf{x}$  by giving more weight to the (training set) vectors that fall closer to  $\mathbf{x}$ ; and third degree splines, a piecewise third degree polynomial model that provides better smoothing and generalization properties than locally weighted ones. For the purpose of comparison, we will also use a baseline model that generates prediction intervals based on a constant variance (homoscedastic) assumption. This is the traditional solution of estimating confidence or prediction intervals from the observed error.

To assess the quality of the results, we will use the measures proposed by Khosravi et al. [8], namely prediction interval coverage probability (PICP), mean PI length (MPIL), normalized mean PI length (NMPIL) and coverage-length-based criterion (CLC). We will analyze the limitations of these measures and propose two additions, namely relative mean PI length (RMPIL) and a revised formulation of coverage-length-based criterion (CLC2).

Thus, the main contribution of this paper is a meta-model that uses quantile regression for creating prediction intervals in real-time. Such model can be easily applied to other contexts and be extended with online learning [11] or stacked regression [12]. In this paper, we focus on the development and validation of such meta-model from a methodological perspective, and leave such extensions for the future work.

To provide a bigger picture for the reader, the next section quickly summarizes the context of this work within our larger project, DynaMIT2.0. It will serve to motivate the remainder of the paper more clearly. We will follow with a literature review (Section III). The methodology will be presented on Section IV, followed by the presentation of our data context and experiments (Section V). We will end this paper with a discussion and conclusions.

## II. CONTEXT

This work belongs to a larger framework called DynaMIT2.0. It corresponds to a next generation dynamic traffic assignment (DTA) real-time traffic prediction system that

inherits most of the aspects of earlier DynaMIT project [3] and adds a few innovative aspects that include new types of data (e.g. feeds from internet, real-time probes, environmental sensors), multi-modality, crisis management advisory, enhanced online and offline calibration and capabilities to better account for uncertainty. The latter is the subject of this research.

There is uncertainty both in the input data as well as in the outputs/predictions. The latter is both due to the inputs themselves but also to the stochasticity and correction of the prediction model. At the input side, DynaMIT2.0 will benefit from information about the quality and reliability of the data; the reliability at the output side is relevant both for the user and for the system itself, particularly for the process of (self) calibration [13].

This article presents a solution that will be capable of participating both at the input and at the output side since it will treat the signal generator as a black box. This signal generator may range from a simple predictor that estimates speed values from a speed/density relationship function (having density as input) to full Origin/Destination travel time prediction. For practical and methodological reasons, the case study of this article will be based on the prediction/estimation of link-based speed. Practically, we have all necessary values (observed densities and speeds) to validate our results and it is in itself a building block for the project. DynaMIT uses speed/density relationship functions [14], [3] to determine vehicle movement in its mesoscopic simulator. These functions are often simplifications of very complex phenomena that may be affected by external factors (e.g. weather status, time of day, incidents) and carry a relevant level of uncertainty, and information on uncertainty can itself be important for DynaMIT’s online calibration process.

Since DynaMIT is not constrained to the speed/density relationship functions, we explore two other distinct solutions, namely LOESS regression and neural networks, that may eventually be integrated in the system.

Methodologically, we prefer to analyze the quality of the model with the inputs first, focusing on a quantity observed in our dataset, speeds, and then leave the outputs for a subsequent step. The latter will include non-observable quantities such as travel times.

## III. LITERATURE REVIEW

The issue of reliability in speed and travel time prediction has received considerable attention in the literature, as it is generally accepted that travelers do not only consider the duration and congestion levels of the trip, but also its certainty, in making pre-trip and en-route choices [15], [16], [17]. Travel time uncertainty causes scheduling costs due to early or late arrival [18].

We focus here on uncertainty for short-term traffic forecasting, therefore we redirect the interested reader to a detailed discussion of uncertainty in medium- and long-term traffic forecasts by De Jong et al. [19].

There are several dimensions to explore in terms of uncertainty treatment within short term traffic prediction. The first one relates to reliability of the input data. Techniques

that somehow do *data fusion* from multiple sensor data are key to consider. They need to take advantage heterogeneity of sensor types (e.g. loop counters and probe vehicles [20]) as well as spatial/temporal correlations. This is a crucial aspect in works like Sun et al.'s where they combine multiple Bayesian networks models [21] into a single one [22] by taking advantage of spatial/temporal correlations between different areas/sensors. In this way, for example in absence of data from a certain sensor, this model is still capable of generating predictions with comparable reliability in a seamless way. Later inspired by this concept, Gao et al. [23] use a graphical lasso (GL) methodology to extract the network correlation structure, as opposed to the earlier used Pearson correlation coefficient from [22]. On a somewhat similar vein to these works, Djuric et al. [24] explore the use of Continuous Conditional Random Fields (CCRF) to speed prediction. CCRF is a probabilistic approach that can incorporate multiple traffic predictors, improve prediction accuracy over regression models, provide information about prediction uncertainty and consider spatial and temporal traffic data correlations.

A second dimension on uncertainty of traffic prediction relates to the model structure itself. Multiple traffic situation scenarios or regimes can occur that may demand different types of models. In this context, solutions that employ ensemble methods have been attracting attention. For example, van Lint et al. [25], [26] presents a framework for real-time, short-term freeway travel time prediction that uses an ensemble of state-space neural networks, which learn to predict travel times directly from data obtained in real-time that also provides confidence estimates which indicate the reliability of the model's outcome, given a certain input. The resulting reliability indicator in this case does not provide upper and lower bounds about the prediction, but instead is a robust indicator of the reliability of that prediction; therefore, it could be used by the operators as a leading indicator of a deterioration of the predictive quality, so that they can seek remedies.

From a Bayesian perspective, an ensemble can be realized as a mixture model: rather than assuming a single underlying distribution, data follows a combination of different distributions. Sun and Xu [27] take this perspective by combining multiple Gaussian Processes (GP) models into a single regression model. A Gaussian Processes model is a kernel-based non-parametric machine learning method with well recognized advantages in terms of adaptability and generalizability, it has been gaining very impressive attention within and beyond the machine learning and data mining communities. It is however very sensitive to training dataset size since it usually needs to memorize all input points. By effectively splitting a hugely complex GP model into a (theoretically infinite) set of smaller GPs, Sun and Xu are able to *divide and conquer* the challenge while at the same time generating an ensemble model that allows multiple individual probabilistic distributions that accommodate to the heterogeneity of traffic characteristics. Due to its probabilistic properties, this model is capable of generating localized confidence intervals.

Another angle on uncertainty treatment relates to the (latent) parameters of the models. It is well known that, particularly

for complex and heterogenous datasets, traditional maximum likelihood approaches often lead to over fitting, and the Bayesian framework is often the proposed solution. In this framework, parameter inference considers not only the training set likelihood but also prior knowledge on the parameters. In such a real-world application, this aspect cannot be ignored. Furthermore, depending on model functional form, such methods are often easily extensible with online learning. For example, Xie et al. [28] apply this framework with Gaussian Processes for short-term traffic flow forecasting. As mentioned above, this model allows for probabilistic interpretation of the model output, but in these cases, however, they keep the assumption of homoscedasticity. Further work exists that equips GPs with heteroscedastic capabilities (e.g. [7]), although to our knowledge not applied to traffic prediction so far.

Still from the Bayesian angle, Fei et al. [29] present a dynamic linear model for online short-term freeway travel time prediction. The prediction result is the posterior travel time distribution that can be employed to generate a single value (typically but not necessarily the mean) travel time as well as a confidence interval representing the uncertainty of travel time prediction. Ghosh et al. [30] use Bayesian time-series models for short-term traffic flow forecasting. Each forecast has a probability density curve with the maximum probable value as the point forecast. Individual probability density curves provide a time-varying prediction interval, unlike the constant prediction interval from classical inference methods.

Regarding the specific estimation of prediction intervals, there has been some relevant work so far. Mazloumi et al. [31] provide a methodology for constructing prediction intervals for neural networks and quantifying the extent that each source of uncertainty contributes to total prediction uncertainty. The authors apply the methodology to bus travel time prediction and obtain quantitative decomposition of the prediction uncertainty into the effect of model structure and inputs data noise. Mazloumi et al. [32] explore the value that traffic flow data can provide to the accuracy of bus travel-time predictions compared with when either temporal variables or scheduled travel times are the base for prediction. While the use of scheduled travel times results in the poorest prediction performance, incorporating traffic flow data yields minor improvements in prediction accuracy compared with when temporal variables are used.

Khosravi et al. [33] present two techniques, (i) delta, based on the interpretation of neural networks as nonlinear regressors, and (ii) Bayesian, for the construction of prediction intervals to account for uncertainties in travel time prediction. The results suggest that the delta technique outperforms the Bayesian technique in terms of narrowness of prediction intervals, while prediction intervals constructed with the Bayesian approach are more robust. Khosravi et al. [34] present a genetic algorithm-based method to automate the process of selecting the optimal neural network model specification. Model selection and parameter adjustments are performed through a minimization of a prediction interval-based cost function, which depends on the properties of the constructed prediction intervals. A review of neural network-based prediction interval methods can be found in [35].

We have been analyzing uncertainty treatment in traffic prediction models that mostly rely on statistical inference or machine learning, but we cannot ignore the entire realm of simulation based approaches, such as dynamic traffic assignment (DTA), such as DynaMIT [3] or DynaSMART [36]. In these cases, the same general questions apply, regarding input data, model parameters or model structure, and some work has been done in studying their sensitivity to uncertainty factors (e.g. [37], [38]). Some of the above approaches can also be applied to this context (e.g. ensembles with different DTA models; Bayesian formulation of input parameters), and in fact such simulation models are themselves capable of *data fusion* in the sense that they can incorporate different types of datasources such as car counters (as volume measurements) and probe vehicle data (as speed/travel time measurements) (see e.g. [38]). Furthermore, by their nature, they bring the extra capability of driving behavior simulation, which can be key to reliable predictions in abnormal situations (e.g. by incorporating driver's reactions to information in incident or crisis scenarios). For example, Antoniou et al. [39] present a framework for the evaluation of the effectiveness of traffic diversion strategies for non-recurrent congestion, which results in travel time savings and increased travel time reliability (in the form of reduced standard deviations of expected travel time when predictive information is provided to the drivers). Waller and Ziliaskopoulos [40] present a chance-constraint system optimum dynamic traffic assignment formulation that can provide solutions with a user specified level of reliability.

To conclude, significant work exists on understanding error structure and level of uncertainty in traffic prediction systems, mostly by incorporating it into the model itself. There is also treatment of uncertainty at the input as well as at the parameters level, to provide interval rather than point estimates. In general, these works are deeply rooted in their traffic prediction model formulations. While this is certainly one of their strengths, it is also a weakness in terms of model extensibility and flexibility.

What we propose here adds the perspective of an agnostic approach, where a meta-model is designed that depends on minimal assumptions and knowledge about the prediction model and its own error structure. This is, to our perception, a novel and relevant contribution to Intelligent Transportation Systems research and practice.

#### IV. METHODOLOGY

Our methodology is designed with the following assumptions in mind:

- There is a *black box* model that continuously generates predictions. These predictions can consist of network performance indicators such as travel times, speeds or densities;
- We focus on the error of such predictions as defined by the (positive or negative) deviations between predicted and observed values. In practice, we transform the error signal into a sequence of deviations  $d = y - \hat{y}$ , where  $y$  is the observed value and  $\hat{y}$  is the value predicted by the black box model;

- The error has a heterogeneous nature with respect to the input dimensions. Although it may generally have a gaussian noise structure, its parameters vary both in terms of mean and variance, i.e., the predictions may not only have heterogenous variance but also be biased;
- Our task is to continuously associate prediction intervals for this error signal, i.e., upper and lower bounds for the predicted value of interest.

##### A. Prediction intervals

We need to distinguish the concept of prediction intervals from that of *confidence intervals*. A 90% confidence interval is expected to contain the population mean in at least 90% of repeated sample experiments while a prediction interval should contain the next predicted value in at least 90% of the times. This subtle difference is discussed throughout the literature (e.g. [41], [42]) and will be relevant to determine the benchmark model that assumes constant variance, explained later in this article.

To obtain the prediction interval bounds, we need to determine a pair of functions,  $d^{\tau-}(\mathbf{x})$  and  $d^{\tau+}(\mathbf{x})$ , that respectively provide the lower and upper quantiles,  $\tau_-$  and  $\tau_+$ , for some  $\tau$  defined by the modeler<sup>1</sup>. This function should give values that underestimate the observed value of  $d$  for only  $(1 - \tau)\%$  of the  $\mathbf{x}$  vectors (and conversely overestimates for  $\tau\%$  of them).

##### B. Quantile regression

Conditional quantile regression provides a solution to this problem. It is a natural option for dealing with heteroscedastic error from a meta-model perspective, as we hope to demonstrate.

We will now explain conditional quantile regression, mostly following earlier work from Roger Koenker [1]. The reader will notice that the major difference to common regression (for the mean value) is the loss function. While common regression uses the sum of squared error, quantile regression applies the tilted loss function.

1) *Quantile regression with linear models*: In contrast to least-squares regression methods, which fit the regression parameters to the conditional mean, quantile regression fits them to the quantiles. The key difference is in the objective function: in quantile regression, rather than minimizing the sum of squared residuals, we use the *tilted loss function*, which is essentially minimal when the exact proportion of values fall below a specific value, the quantile. If we are looking for the quantile  $\tau$ , the loss of choosing  $d^\tau$  is defined as  $\rho_\tau(u) = u(\tau - I(u < 0))$ , with  $u = d - d^\tau$  and  $I(\cdot)$  as an indicator function. In Figure 1, we show the general intuition. Notice that, for  $\tau = 0.5$ , the loss is evenly distributed throughout both sides, i.e., the ideal value falls exactly in the middle (the median) of the  $u$  line. On the other hand, for  $\tau = 0.75$ , there is an imbalance (the *tilt*) where staying below the chosen value is less costly.

Another way to see this function is the following:

<sup>1</sup>To keep notation uncluttered, instead of  $d^\tau(\mathbf{x})$ , we will use the following notation:  $d^\tau$  corresponds to the conditional quantile function while  $d_i^\tau$  is the value of that function for input vector  $\mathbf{x}_i$

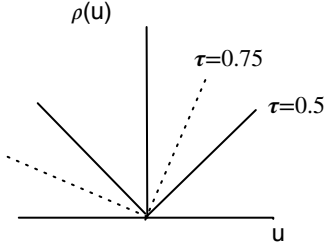


Fig. 1. Tilted loss function,  $\rho(u)$ .

$$\rho_\tau(d_i - d^\tau) = \begin{cases} \tau(d^\tau - d_i) & d_i \geq d^\tau \\ (1 - \tau)(d_i - d^\tau) & d_i < d^\tau \end{cases} \quad (1)$$

Thus, we can determine the expected loss,  $E_{\rho_\tau}(D - d^\tau)$ , of choosing  $d^\tau$  to be the  $\tau$  quantile for random variable  $D$ , given a set of samples  $d_i$ , with  $i = 1..N$ :

$$(\tau - 1) \sum_{d_i < d^\tau} (d_i - d^\tau) + \tau \sum_{d_i \geq d^\tau} (d^\tau - d_i) \quad (2)$$

It will be rare to have a set of (training) samples,  $d_i$ , for each input vector  $\mathbf{x}$  such that we can precisely estimate its *local* quantile in this way. Instead, we can use the distribution of values throughout an entire dataset to estimate a single function such that  $d_i^\tau = \mathbf{x}_i^T \boldsymbol{\beta}_\tau$ . I.e., at each input vector  $\mathbf{x}_i$ , there will be a potentially unique quantile value,  $d_i^\tau$ , that is a function of its components. Thus, for conditional quantile regression, we need to minimize

$$(\tau - 1) \sum_{\substack{(\mathbf{x}_i, d_i) \in S \\ d_i < \mathbf{x}_i^T \boldsymbol{\beta}_\tau}} (d_i - \mathbf{x}_i^T \boldsymbol{\beta}_\tau) + \tau \sum_{\substack{(\mathbf{x}_i, d_i) \in S \\ d_i \geq \mathbf{x}_i^T \boldsymbol{\beta}_\tau}} (\mathbf{x}_i^T \boldsymbol{\beta}_\tau - d_i) \quad (3)$$

with respect to the parameters  $\boldsymbol{\beta}_\tau$  throughout the entire dataset  $S = (\mathbf{X}, \mathbf{d})$  of size  $N$ , where for each input vector  $\mathbf{x}_i \in \mathbf{X}$ , we have the corresponding observed deviation value,  $d_i \in \mathbf{d}$ . In other words, we need to estimate the values for  $\boldsymbol{\beta}_\tau$  since  $S$  and  $\tau$  are all given. This can be reformulated as a linear programming problem, solvable by the simplex method. We will not add further details on this procedure here, so we redirect the interested reader to Koenker's book [1] on the subject.

For the purpose of illustration, we created two toy models of the form  $y = ax + b + \epsilon$  where  $\epsilon$  is distributed either as  $N(0, x)$  (variance grows with  $x$ ) or  $N(0, gp(0, 1))$  (variance is sampled from a Gaussian Processes prior). We call them noise models 1 and 2, respectively. We assume that our (black box) predictor,  $\hat{y}$ , corresponds to the denoised model,  $\hat{y} = ax + b$ . We then calculated the set of deviations,  $d_i = y_i - \hat{y}_i$ , and applied the conditional quantile regression method just described. The bounds become  $\hat{y}_i^{\tau-} = \hat{y}_i + d_i^{\tau-}$  and  $\hat{y}_i^{\tau+} = \hat{y}_i + d_i^{\tau+}$ , respectively for lower and upper bounds at each  $\mathbf{x}_i$ . Plots in Figure 2 show this data together with the 5% and 95% quantile bounds given by our model. Notice that, in this case,  $x$  or  $y$  do not correspond to any real world phenomena, they are purely chosen for demonstrative purposes.

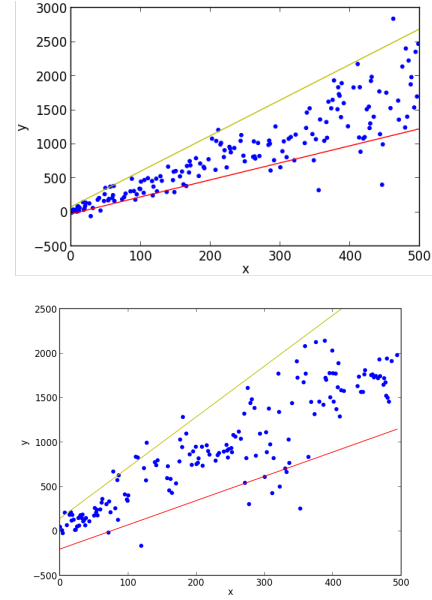


Fig. 2. Toy model 1 (top) and model 2 (bottom). We show the underlying predictions (of  $y$  as a function of  $x$ ) as well as the lower and upper bounds.

This procedure yields a linear model for each  $\tau$  quantile, thus it is insufficient when the error varies non-linearly throughout the entire dataset. A solution is to estimate an independent model for each prediction point by applying locally weighted regression, essentially assigning more weight to neighboring vectors according to some distance function, for example the *squared exponential*,  $s_{i,j} = e^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2h^2}}$ , where  $h$  is known as the bandwidth, and  $\mathbf{x}_i$  and  $\mathbf{x}_j$  are any two vectors in the dataset. Thus, we rewrite equation 3 into:

$$(\tau - 1) \sum_{\substack{i: (\mathbf{x}_i, d_i) \in S \\ d_i < \mathbf{x}_i^T \boldsymbol{\beta}_\tau}} s_{i,0} (d_i - \mathbf{x}_i^T \boldsymbol{\beta}_\tau) + \tau \sum_{\substack{i: (\mathbf{x}_i, d_i) \in S \\ d_i \geq \mathbf{x}_i^T \boldsymbol{\beta}_\tau}} s_{i,0} (\mathbf{x}_i^T \boldsymbol{\beta}_\tau - d_i) \quad (4)$$

where  $s_{i,0}$  represents the distance between the prediction point and vector  $\mathbf{x}_i$  from the dataset. Figure 3 illustrates the effect. The wiggly behavior is due to the preference to the local versus global fits. We will discuss this in the next section.

Locally weighted regression has been used with success in traffic estimation and prediction (e.g. [10], [43]), and falls in the family of vast literature on non-parametric models such as gaussian processes [44], k-nearest neighbor regression [4] or support vector regression [45].

2) *Quantile regression with spline models*: In our context, locally linear models have three relevant problems. We need to estimate multiple individual models; such models are limited to a linear form and as a result the overall (piece-wise linear) function is not smooth; unless the point's neighborhood is representative enough, it is more prone to overfitting than a global model. The two latter points are clearly visible in Figure 3.

A solution to these problems proposed in [1] is to use a splines function. This function is decomposed into a sequence of piece-wise polynomial functions that are linked together at

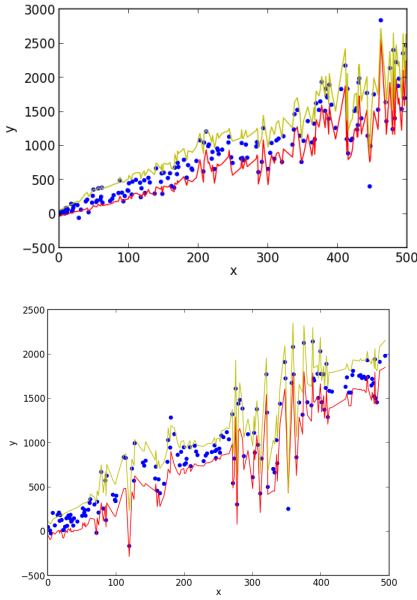


Fig. 3. Locally linear models on toy models 1 (top) and 2 (bottom).

$M$  knot points. The splines function needs to be differentiable in all its points, including at the knots. We apply a well known decomposition process, the B-splines, where the spline function is expressed as a linear combination of *Basis* splines with degree  $k$ . So now we have a new definition for  $d_i^\tau$  that adds a splines component to the linear combination  $\mathbf{x}_i^T \beta_\tau$  defined earlier. Formally,

$$d_i^\tau = \mathbf{x}_i^1{}^T \beta_\tau^l + \sum_{m=-1}^{M+1} a_m B_m(\mathbf{x}_i^s) \quad (5)$$

where  $\mathbf{x}_i^1$  and  $\beta_\tau^l$  correspond to the input features and parameters that remain in the linear part of the model, and  $\mathbf{x}_i^s$  corresponds to the input features that are applied splines. There is an  $a_m$  coefficient for each of the  $M + 2$  cubic B-spline functions,  $B_m$ . Each one of these functions calculates the contribution for  $\mathbf{x}_i^s$  of the piece-wise cubic function between knots  $m$  and  $m+1$ . The B-spline functions,  $B_m$ , are the subject of extensive literature and are implemented in a wide variety of software packages. For further details, we redirect the reader to [46].

To estimate and run this model, we used the packages *splines* and *quantreg* [47], available in *R* [48], together with the objective function specified by equation 3. The most common form of spline models uses a  $3^{rd}$  degree polynomial, the *cubic splines*. Unless otherwise mentioned, we will keep this choice.

From the point of view of quantile regression, cubic splines allow for capturing local effects while still being differentiable at every point, hence the smoothing effect and better generalization properties.

In Figure 4, we illustrate the result with the toy model above described. In this case, we use  $d_i^\tau = \sum_{m=-1}^{M+1} a_m B_m(x)$ .

We will not provide further details on quantile regression so we redirect the interested reader to fundamental literature,

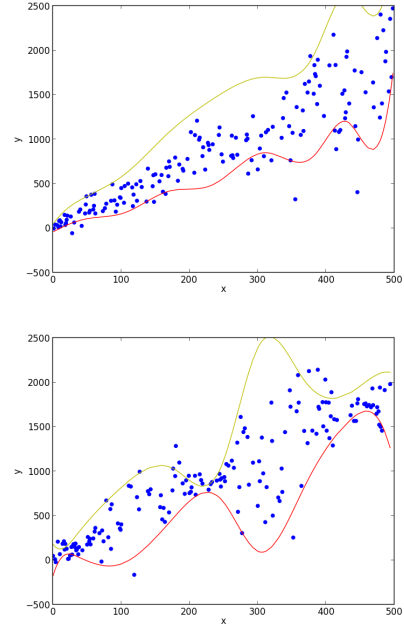


Fig. 4. Cubic splines on toy models 1 (top) and 2 (bottom).

namely Koenker's book [1], the gentle introduction of Cade and Noon [9], or specific contributions on local linear models [49] or application with gaussian processes [44]. We also advise the use of the very well designed and complete quantile regression package, *quantreg*, available for *R* [47] for those keen to implement this method.

#### C. Prediction intervals for homoscedastic processes

As mentioned earlier, we will also use a baseline model where we use a known constant (homoscedastic) variance to generate the interval bounds for each prediction,  $\hat{y}$ . First, we obtain the observed error variance from the training set,  $\sigma^2$ , then we determine the prediction interval bounds,  $\hat{y}^{\tau-}$  and  $\hat{y}^{\tau+}$ , according to

$$\hat{y}^{\tau\pm} = \hat{y} \pm t_p \sigma \sqrt{1 + 1/n}, \quad (6)$$

where  $t_p$  corresponds to the quantile  $100((1 - p)/2)\%$  of Student's t-distribution with  $n-1$  degrees of freedom. Such procedure is well know and a common practice (e.g. [50]).

#### D. Putting it all together

The quantile regression methods presented provide a function  $d^\tau$  that, for any given input vector, yields the quantile at the  $\tau$  level. Our method is to use 2 independent models, one for predicting the lower bound ( $\tau_-$ ), the other for the upper bound ( $\tau_+$ ), that will be estimated from a time series of error deviations. For the input vector,  $\mathbf{x}$ , we will consider the black box predicted value,  $\hat{y}$ , the time of day (*peak, non peak*), as well as the  $L$  auto-regression lags, i.e.  $L$  previous values of error. A preliminary autocorrelation analysis revealed that  $L = 3$  should be an acceptable value throughout the dataset, i.e. in general, up to the third lag, the signal is autocorrelated

by 0.5 or above. Other input features could be added trivially, such as weather status or incident information, however we do not have such information for the available datasets.

The target value is, as defined before, given by  $d_i = y_i - \hat{y}_i$ , the deviation between the predicted and observed value for every single input vector  $\mathbf{x}_i$ . Our algorithms aim to predict the upper and lower bound values,  $\hat{y}_i^{\tau^+}$  and  $\hat{y}_i^{\tau^-}$ , resp., such that, in  $100 * (\tau^+ - \tau^-)\%$  of the times, the target value falls within the interval. We will choose  $\tau^+$  and  $\tau^-$  to be 0.95 and 0.05, respectively. This implies that, for the constant variance model, we assign  $t_p = 1.645$  (the two-sided 90% value for the t-distribution).

### E. Performance measures

While for pointwise regression the performance measures are very well defined and accepted (e.g. root mean squared error, correlation coefficient, relative squared error), the same cannot be said for evaluating interval regression. This happens because we need to consider both whether the predicted intervals did contain the target value as well as how precise their bounds were. In fact, if we assign extreme values to the bounds, we trivially get an interval predictor that contains the targets for 100% of the times. Conversely, we can also get very narrow intervals that repeatedly fail. A few solutions were proposed by Khosravi et al. [8]:

- Prediction interval coverage probability (PICP),

$$PICP = \frac{1}{n} \sum_{i=1}^n c_i,$$

where  $c_i = 1$  if  $y_i \in [\hat{y}_i^{\tau^-}, \hat{y}_i^{\tau^+}]$

- Mean prediction interval length (MPIL),

$$MPIL = \frac{1}{n} \sum_{i=1}^n (\hat{y}_i^{\tau^+} - \hat{y}_i^{\tau^-})$$

- Normalized MPIL,  $NMPIL = \frac{MPIL}{R}$ , with  $R$  being defined *a priori*
- Coverage-length-based criterion (CLC),

$$CLC = NMPIL(1 + e^{(-\eta(PICP - \mu))})$$

with  $\eta$  and  $\mu$  as two controlling parameters.

Ideally,  $PICP$  should be as close as possible to the  $(\tau^+ - \tau^-)$ . In our case, this should be 0.9. For  $MPIL$ , the lowest value possible is sought.  $NMPIL$  demands a specific parameter,  $R$ , that is not clearly defined in [8], so we propose a specific implementation, the relative mean prediction interval length ( $RMPIL$ ),

$$RMPIL = \frac{1}{n} \sum_{i=1}^n \frac{(\hat{y}_i^{\tau^+} - \hat{y}_i^{\tau^-})}{|y_i - \hat{y}_i|}$$

i.e., we normalize the interval length by the observed error. Large intervals are in fact *necessary* whenever actual error

is large itself, otherwise it is not possible to cover the target value.

In the  $CLC$  measure, the  $\mu$  parameter has a very clear meaning: it represents the desired  $PICP$  value. On the other hand, the parameter  $\eta$  is difficult to materialize; it magnifies the *importance* of reaching the desired  $PICP$ . Furthermore, the interval length measure,  $NMPIL$ , has a potentially negligible role. If we have a  $PICP$  below  $\mu$ , even by a very small amount,  $CLC$  becomes unreasonably affected. For example, with  $\mu = 0.9$ ,  $\eta = 200$  as in [8], a model with coverage very close to the ideal such as 89% ( $PICP = 0.89$ ) would obtain a  $CLC$  higher than  $NMPIL^2$ , in fact being worse than any other model with  $PICP$  of 0.9 or above that relaxes the bounds up to  $NMPIL^2$ . In other words, unreasonably large intervals would be preferable than a more precise model with  $PICP$  marginally below the objective.

The value of  $\eta$  serves to control this effect, however its choice is arbitrary and becomes a challenge in itself. Assigning a very low value would also underestimate the importance of  $PICP$ . After repeated empirical tests, we decided to set the value to 100. We also propose a simpler measure, called  $CLC2$ :

$$CLC2 = e^{(-RMPIL(PICP - \mu))}$$

This measure gives both  $MPIL$  and  $PICP$  comparable roles and removes the subjective control parameter. We will use all defined measures just discussed in our evaluation.

## V. EXPERIMENTS

### A. Experimental design

Two freeway data-sets from Irvine, CA, and Tel Aviv in Israel have been used for this research. In both cases, weekday data were used. The Irvine data set includes five days of sensor data from freeway I-405. The application involved training/calibration with four days of data and subsequent testing/validation of the model framework for the fifth day (not used in the calibration). Data from 10am to 12midnight have been used, since this period includes the (pm) peak flow for this direction. Speed, occupancy and flow data over 2-minute intervals were available for calibration and validation. Occupancy data have been converted to density using a relationship from May ([14], eq. 7.2 in p. 193).

The second data set was collected in Highway 20 (Ayalon Highway), a major intra-city freeway running through the center of Tel Aviv in Israel. Four days of data were used for the training of the models and a different fifth day was used for validation. Speed, occupancy and flow data were available and were aggregated over 5-minute intervals. Occupancy data have been converted to density using the same relationship as above.

The following different cases are developed, based on the type of approach that is used for state (where applicable) and speed prediction:

- 1) Typical speed-density relationship: a commonly used relationship is fit to the speed and density data of the training data set.

The following speed-density relationship model was used as the reference model [3]:

$$u = u_f \left[ 1 - \left( \frac{\max(0, k - k_{min})}{k_{jam}} \right)^\beta \right]^\alpha \quad (7)$$

where,  $u$  denotes the space mean speed,  $u_f$  the free flow speed,  $k$  the density,  $k_{min}$  the minimum density,  $k_{jam}$  the jam density, and  $\alpha$  and  $\beta$  are model parameters. This is a variant of the speed-density traffic flow theory relationship that is commonly used in mesoscopic traffic simulation models. For example, this is the relationship used in the DynaMIT model [3] and very similar to the relationship used in the DynaSMART [36] and mezzo [51] models.

The estimated relationship is then used to calculate speed values based on the densities in the test data set. The true densities (instead of predicted) are used in this process, thus eliminating any prediction error and providing an even better than expected prediction of speeds for this baseline model.

- 2) Speed prediction framework presented in Antoniou et al. [10]: a complete state and speed prediction framework has been applied using the available data. The methodology comprises training and application steps. During the training step archived surveillance data are used to (A) identify the various traffic states through clustering the available observations; (B) estimate the transition processes between these regimes; and (C) estimate cluster-specific traffic models. This information is stored into a knowledge base and supports the application of the framework. As new measurements become available, they are (D) classified into the appropriate regimes and, based on the transition processes and the short-term evolution of the traffic state, (E) short-term predictions of the traffic state are performed using the applicable estimated transition processes. Furthermore, (F) the appropriate flexible traffic model is retrieved and applied to the new observations to (G) perform speed predictions. In this case, the optimal number of states is used, i.e. the number of states that minimizes the Bayesian Information Criterion (BIC), based on the results from the model-based clustering algorithm.
- 3) Simplified framework presented in Antoniou et al. [10]: The complete state and speed prediction framework is used, but neural networks are used for the clustering and classification steps. This is a simpler approach that is implemented in order to assess the incremental benefits of the proposed framework components.

These three approaches will be our *black box* speed prediction models. Henceforth, for simplicity of reference, we will call the first approach as *spddsty*; the second one will be LOESS (locally weighted scatterplot smoothing) and the last one will be NNNet.

For each dataset and each black box model, we will run our three prediction interval meta-models: with constant variance (*const*); splines quantile regression (*splines*); and locally weighted quantile regression (*local*). We will use the first 2/3 of the dataset (ordered in time) for training and 1/3 for testing.

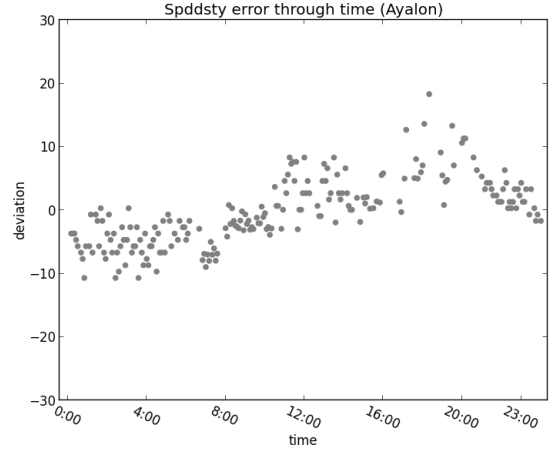


Fig. 5. Speed-density function performance in Ayalon.

## B. Results

Table I shows the overall results obtained. Unsurprisingly, we can see that the constant variance model achieves PICP values close to the intended prediction interval coverage (of 0.9) in all cases but it does so at the expense of the largest interval ranges. This also indicates that the variance observed in the training set is similar to that of the test set, i.e. they seem to be adequately balanced. Except for two models in Ayalon (NNet and Loess), the quantile regression models obtain performance with good PICP, interval length and CLC measures. And we can also observe that the CLC2 measure provides a more balanced evaluation with regards to PICP and interval length.

TABLE I  
OVERALL RESULTS TABLE.

blckbx model	dataset	meta model	PICP	MPIL	RMPIL	CLC	CLC2
Loess	Irvine	local	0.82	8.79	10.86	6.70E+04	2.39
Loess	Irvine	const	<b>0.88</b>	9.69	12.50	<b>1.71E+03</b>	1.33
Loess	Irvine	splines	0.81	<b>8.67</b>	<b>10.76</b>	8.09E+05	<b>1.21</b>
Loess	Ayalon	local	0.59	<b>4.52</b>	<b>5.75</b>	2.86E+14	5.90
Loess	Ayalon	const	<b>0.85</b>	8.83	11.22	<b>3.07E+04</b>	<b>1.85</b>
Loess	Ayalon	splines	0.69	7.42	9.80	1.25E+11	6.70
Spddsty	Irvine	local	0.74	<b>15.04</b>	<b>12.83</b>	1.47E+08	7.36
Spddsty	Irvine	const	0.92	20.88	18.12	<b>3.56E+02</b>	0.638
Spddsty	Irvine	splines	<b>0.95</b>	21.59	18.95	3.62E+02	<b>0.408</b>
Spddsty	Ayalon	local	0.74	<b>8.05</b>	<b>9.64</b>	2.00E+08	4.75
Spddsty	Ayalon	const	0.89	17.50	19.00	1.14E+03	1.16
Spddsty	Ayalon	splines	<b>0.93</b>	12.42	12.15	<b>1.55E+02</b>	<b>0.698</b>
NNet	Irvine	local	0.87	<b>9.50</b>	<b>17.17</b>	<b>8.08E+02</b>	<b>1.51</b>
NNet	Irvine	const	<b>0.88</b>	9.56	18.07	3.57E+03	1.71
NNet	Irvine	splines	<b>0.88</b>	10.05	19.33	4.08E+03	1.56
NNet	Ayalon	local	0.66	<b>4.44</b>	<b>15.43</b>	<b>6.71E+11</b>	<b>5.61</b>
NNet	Ayalon	const	<b>0.86</b>	9.53	42.22	1.08E+05	39.4
NNet	Ayalon	splines	0.69	6.46	23.75	7.34E+11	146

Before getting into further details, let us take a quick look at a few of the error signals,  $d_i$ . Figures 5, 6, 7 and 8 show a few examples. Due to space constraints, we focus on the more challenging case of Ayalon, adding an example from Irvine.

There is a striking difference in variance between the speed/density function and the other two models. And it could be argued that this variance changes in time (x axis)



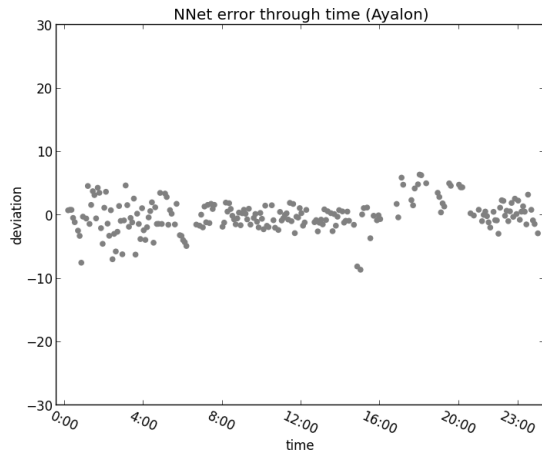


Fig. 6. Neural network performance in Ayalon.

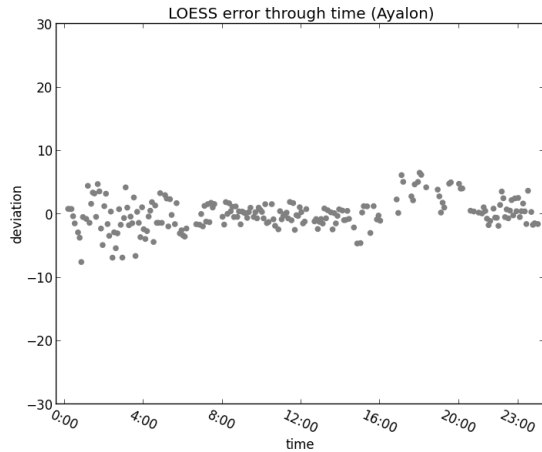


Fig. 7. LOESS performance in Ayalon.

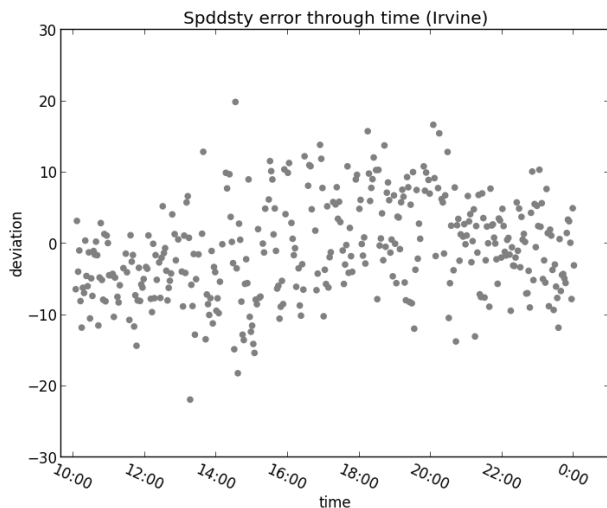


Fig. 8. Speed-density function performance in Irvine.

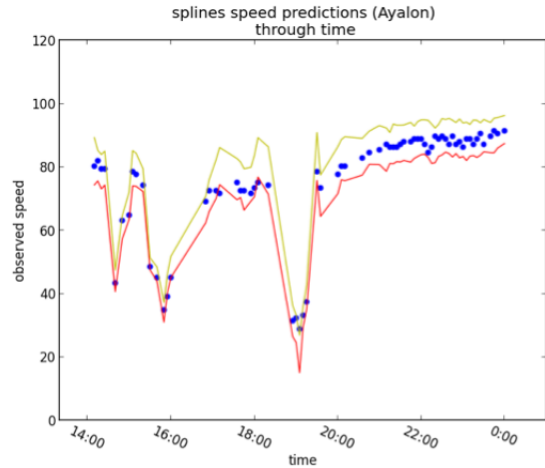


Fig. 9. Prediction intervals for Ayalon speed density model based on splines.

in all cases, particularly for the speed/density function. For the Irvine case, we can say that the same behavior applies, although less sharply. Recalling Table I, the obvious interpretation is that quantile regression is a good option for cases with high, heteroscedastic, variance but, under lower and less varied error, the traditional confidence/prediction intervals approach may be ideal if we value coverage (*PICP*) more than interval length (*RMPIL*).

To better illustrate the actual performance of the algorithms, we plot the prediction intervals through time between the splines quantile regression and constant variance models, for two black boxes: speed density (Figures 9 and 10) and LOESS (Figures 11 and 12). Obviously, constant variance model always generates the same interval length, so its coverage depends essentially on the bias of the model. The speed density function model seems more biased but together with the splines model it is clearly capable obtaining the best performance. If we take a look at CLC2, this same conclusion would apply to the Irvine case, although it is arguable that the splines quantile regression obtained unnecessarily high *PICP* at the expense of intervals that are much larger than necessary. Figure 13 illustrates this point.

## VI. DISCUSSION

In this article, we demonstrate a methodology that is capable of obtaining prediction interval bounds for a black box predictor that depends on very few weak assumptions. These are: the error quantile bounds of the target variable can be defined as a function of some available input vector  $\mathbf{x}$ ; this function has either a linear or polynomial form (potentially composed into a global splines model); and the training and test sets should be independently and identically distributed (iid). There are no other constraints with respect to the error form or the properties of the black box model.

Of course, this versatility does not come without a cost: being disconnected from the source itself (the black box model) implies that we are treating the effects, not the cause. If the traffic prediction engine has very poor performance, our

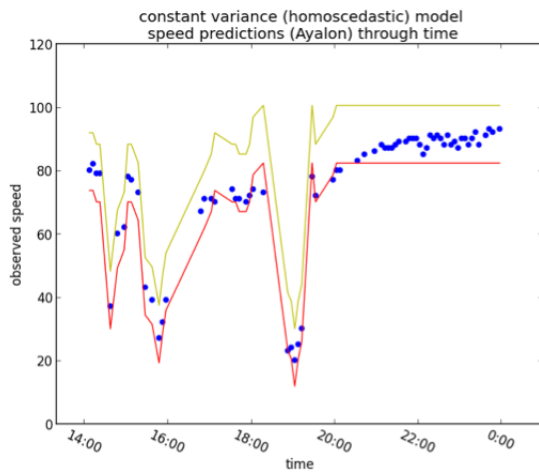


Fig. 10. Prediction intervals for Ayalon speed density model based on constant variance.

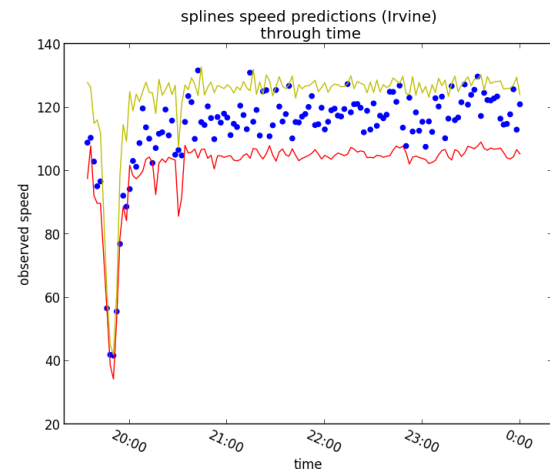


Fig. 13. Prediction intervals for Irvine spddsty model based on splines.

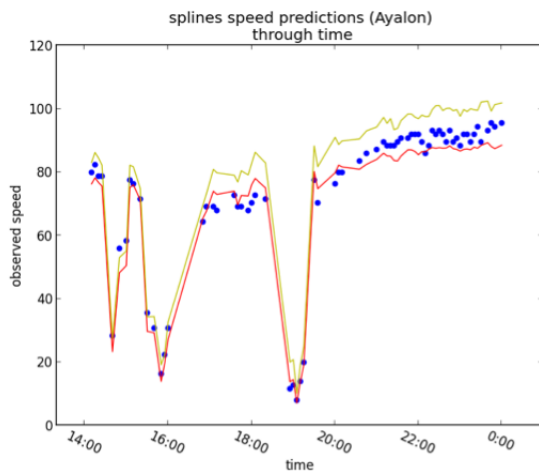


Fig. 11. Prediction intervals for Ayalon LOESS model based on splines.

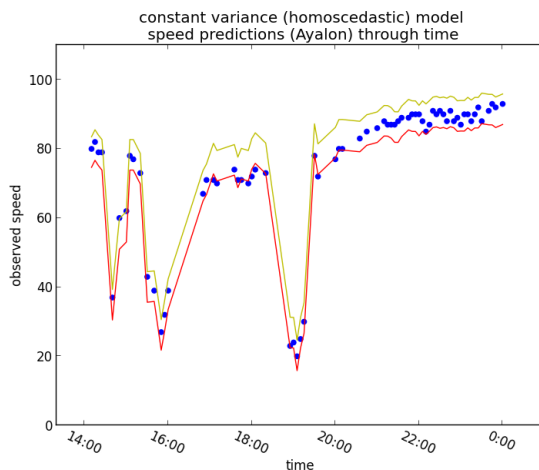


Fig. 12. Prediction intervals for Ayalon LOESS model based on constant variance.

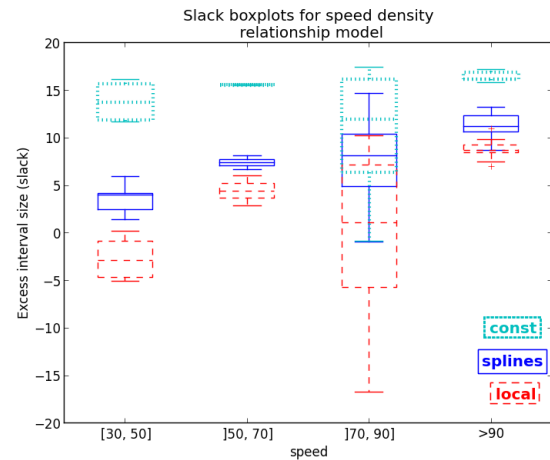


Fig. 14. Slack box plots for Ayalon spddsty model.

providing very large bounds.

We could see that, from the experiments, it is not trivial to evaluate the quality of our three models independently of the quality of the underlying predictions. For example, if the observed speed value is 50km/h and the predictor gives 120 km/h, then the ideal meta-model should only allow enough *slack* to cover these two values<sup>2</sup>. In this case, the smallest slack would be  $120 - 50 = 80$  and any higher value would be erring on the conservative side, while a lower value would be failing to cover the observed value. The *RMPIL* was designed to capture this effect, but we can better see how the several algorithms behave in terms of slack amounts with a visualization. Figures 14 and 15 present the box plots for the speed density function and LOESS for Ayalon, respectively. We show the (extra) slack performance for each of the three meta-models. On the x axis, we show the observed speed.

In practice, a box plot for a good model should be centered slightly above 0, preferably enough not to go below that

<sup>2</sup>We are assuming that the predicted value should be contained in the interval. In our experiments, this was generally the case but the models could learn otherwise.

algorithm won't do more than emphasizing it more clearly, by

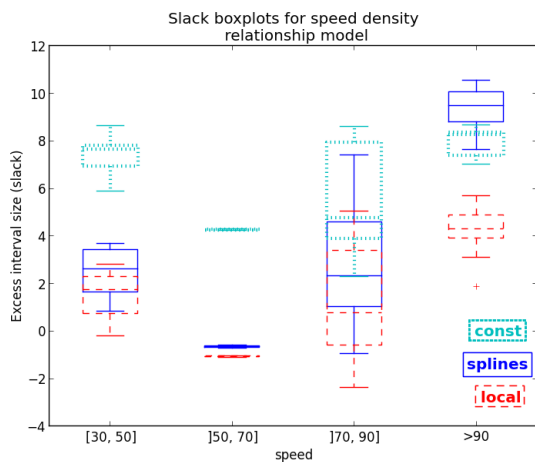


Fig. 15. Slack box plots for Ayalon LOESS model.

value. I.e., it would provide just enough slack above the ideal interval length, to err on the conservative side. Except for a few exceptions, we can clearly see that the constant variance model chooses the largest intervals. The splines model seems to be better than the local one in the sense that the latter obtains more negative slacks. This is coherent with what we see in Figures 3 and 4 of our toy model and with the related discussion on over fitting and smoothness.

Let us now summarize a few other observations from the experiments:

- The quantile models rarely underestimate the upper bound but sometimes overestimate the lower bound. This can be concerning in the case of speeds since it corresponds to an optimistic perspective. For example, DynaMIT2.0 may fail to anticipate congestions properly;
- As a consequence of the above, the *PICP* in some of the quantile models is very low, particularly for Ayalon NNet and Loess. The constant variance model definitely works better in those cases but we note that it is at the expense of large intervals (in the NNet, the *RMPIL* is 42.22);
- After testing with changing the splines degree to other values  $\{1,2,4,5\}$ , the performance degrades;
- After testing different bandwidths,  $h$ , for the locally weighted quantile regression  $\{5, 10, 50, 100, 200\}$ , the value that provides best results, used throughout the experiments, is 100. This is the best balance between overfitting (small values) and noise (high values) for this specific dataset, but we cannot speculate further about its meaning.
- In terms of computational performance, these algorithms spent negligible time in training and prediction on our dataset. However, the case study deals with a single sensor for each case study and a dense network coverage might lead to non-negligible run times. It is expectable that the locally weighted regression model degrades fast with dataset size because it effectively estimates and predicts a new model for each query. The splines model will be slower to estimate but this can be done on an off-

line basis. The constant variance model obviously poses minimal constraints to dataset size and can be estimated on an off-line basis.

An important issue may arise when using conditional quantile regression: the problem of *crossed quantiles*, when, for a given  $x_i$ , the upper quantile is below the lower quantile. We were minimally affected by this phenomenon so it was neglected in the work presented. However, it deserves a more cautious treatment in practice, particularly in light of available literature on the subject (e.g. [52], [53]).

Another concern relates to the sensitivity of the model to traffic state dynamics. We want a model that, under stress conditions, is able to keep the expected reliability. While it is not easy to replicate or identify such conditions in our dataset, we can get some insights by comparing the performance for peak/off-peak periods. Starting with the splines model, we see a somewhat surprising result: during peak periods, its *PICP* actually increases in all cases (e.g. it improves from 0.87 to 0.91 in Irvine NNet). This obviously happens at the expense of larger prediction intervals. Our intuition is that, given similar circumstances in the training set (higher local variance), the quantiles will tend to err on the safe side. For the homoscedastic model, the opposite happens (e.g. it degrades from 0.89 to 0.55 in Irvine NNet), which is not surprising given the total lack of flexibility of this model. Finally, the locally weighted quantile regression model presents mixed results (e.g. it improves from 0.73 to 0.93 in Irvine NNet, while it degrades from 0.88 to 0.81 in Irvine Loess). We should however be cautious in this analysis despite these results being encouraging. Traffic state dynamics will probably be more challenging under exceptional circumstances (e.g. incidents, special events, harsh weather) than the duality of peak/off-peak. Further work will need to be done at this respect.

Finally, these experiments intended to test our methodology without pretending an immediate real-world application with these black box models. Although relevant for this trial, the dataset is somewhat poor in contextual detail that should actually improve the prediction interval capabilities. For example, having information on incidents, weather or special events would certainly lead to a more meaningful real-world application.

## VII. CONCLUSIONS AND FUTURE WORK

We introduced a meta-model methodology that is capable of estimating, in real-time, the prediction intervals for a traffic prediction system by analyzing its error history. It does so by applying conditional quantile regression to the time series signal of the observed error and also considering the predicted value itself. The lower and upper bounds become simply the lower and upper quantiles, as per choice of the modeler. We implemented this concept and tested it with two case-studies in Israel and California, USA. In each case-study, the original prediction task consisted of predicting speeds from sensed densities.

We compared our prediction interval methodology with a constant variance baseline and the experiments show that quantile regression approaches are capable of determining

smaller prediction intervals and comparable accuracy (or prediction interval coverage probability, *PICP*) in most cases. More importantly, under high variance and heteroscedasticity, our approach is more adequate than using traditional confidence intervals. This is particularly relevant when the error structure is not known.

We specifically applied this concept to speed prediction but it can be tested with any other regression task with very little effort. For example, if we want to add interval prediction capabilities to an existing bus arrival prediction algorithm, we only need access to the feed of predictions and observations as well as any other variable that is believed to correlate to the prediction error (e.g. traffic conditions, weather status). The model implementation and experimental methodologies are essentially the same as in this paper.

Future research includes an investigation into the role of different network geometry and other characteristics into the performance of the presented approach, as well as the impact of different prediction intervals. In this work, a preliminary assessment of the robustness of the approach was achieved by looking at data from different networks (i.e. Israel vs. California), and different prediction intervals (i.e. 2 vs 5 min).

We also introduced two new performance measures that aim to better evaluate prediction interval models, both extending the previous work of Khosravi et al. [8] aiming for a better balance between *PICP* and interval length. We add the notion of *slack* as the ideal interval length corresponding to the difference between observed and predicted.

The next step consists of the extension of this work with the gaussian processes algorithm, from a Bayesian approach perspective, building on the work from Boukouvalas et al. [44] and Lázaro-Gredilla et al. [7]. We expect a performance at least comparable to the splines model, while becoming a very good solution for online learning. Any real-world implementation of a prediction intervals meta-model needs to accompany performance evolution of its associated prediction model (online learning).

This project is integrated in the DynaMIT2.0 framework, and participates both as an input data quality estimator as well as a service for the end user, providing upper and lower bounds for predictions of traffic information such as travel times, speeds or volumes.

#### ACKNOWLEDGMENTS

The authors would like to thank Prof. Wilfred Recker from University of California, Irvine, and Dr. Tomer Toledo from Technion Institute of Technology, Israel, for providing the data used in this research. The authors would like to thank the anonymous reviewers for their insightful comments that strongly contributed to the improvement of this article.

This research was supported by the National Research Foundation Singapore through the Singapore MIT Alliance for Research and Technology's FM IRG research programme.

#### REFERENCES

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

- [2] M. Ben-Akiva, M. Bierlaire, D. Burton, H. Koutsopoulos, and R. Mishalani, "Network state estimation and prediction for real-time traffic management," *Networks and Spatial Economics*, vol. 1, no. 3-4, pp. 293–318, 2001.
- [3] M. Ben-Akiva, H. N. Koutsopoulos, C. Antoniou, and R. Balakrishna, "Traffic simulation with dynamit," in *Fundamentals of Traffic Simulation*, J. Barceló, Ed. New York, NY, USA: Springer, 2010, pp. 363–398.
- [4] P. M. Robinson, "Asymptotically efficient estimation in the presence of heteroskedasticity of unknown form," *Econometrica*, vol. 55, no. 4, pp. 875–891, 1987.
- [5] S. Chib and E. Greenberg, "On conditional variance estimation in nonparametric regression," *Statistics and Computing*, vol. 23, pp. 261–270, 2013.
- [6] H. White, "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity," *Econometrica*, vol. 48, pp. 817–838, 1980.
- [7] M. Lázaro-gredilla and M. K. Titsias, "Variational heteroscedastic gaussian process regression," in *In 28th International Conference on Machine Learning (ICML-11)*. ACM, 2011, pp. 841–848.
- [8] A. Khosravi, E. Mazloumi, S. Nahavandi, D. Creighton, and J. W. C. Van Lint, "Prediction intervals to account for uncertainties in travel time prediction," *Intelligent Transportation Systems, IEEE Transactions on*, vol. 12, no. 2, pp. 537–547, 2011.
- [9] B. S. Cade and B. R. Noon, "A gentle introduction to quantile regression for ecologists," *Frontiers in Ecology and the Environment*, vol. 1, no. 8, pp. 412–420, 2003.
- [10] C. Antoniou, H. N. Koutsopoulos, and G. Yannis, "Dynamic data-driven local traffic state estimation and prediction," *Transportation Research Part C: Emerging Technologies*, vol. 34, pp. 89 – 107, 2013. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0968090X13001137>
- [11] S. Shalev-Shwartz, "Online Learning: Theory, Algorithms, and Applications," Ph.D. dissertation, The Hebrew University of Jerusalem, Jul. 2007.
- [12] M. Neumann, K. Kersting, Z. Xu, and D. Schulz, "Stacked gaussian process learning," in *Data Mining, 2009. ICDM '09. Ninth IEEE International Conference on*, 2009, pp. 387–396.
- [13] C. Antoniou, M. Ben-Akiva, and H. N. Koutsopoulos, "Nonlinear kalman filtering algorithms for on-line calibration of dynamic traffic assignment models," *Intelligent Transportation Systems, IEEE Transactions on*, vol. 8, no. 4, pp. 661–670, 2007.
- [14] A. May, *Traffic Flow Fundamentals*. New Jersey: Prentice Hall, 1990.
- [15] M. A. Abdel-Aty, R. Kitamura, and P. P. Jovanis, "Using stated preference data for studying the effect of advanced traffic information on drivers' route choice," *Transportation Research Part C: Emerging Technologies*, vol. 5, no. 1, pp. 39–50, 1997.
- [16] H. S. Mahmassani and Y.-H. Liu, "Dynamics of commuting decision behaviour under advanced traveller information systems," *Transportation Research Part C: Emerging Technologies*, vol. 7, no. 2, pp. 91–107, 1999.
- [17] T. Toledo and R. Beinaker, "Evaluation of the potential benefits of advanced traveler information systems," *Journal of Intelligent Transportation Systems*, vol. 10, no. 4, pp. 173–183, 2006.
- [18] D. Ettema and H. Timmermans, "Costs of travel time uncertainty and benefits of travel time information: conceptual model and numerical examples," *Transportation Research Part C: Emerging Technologies*, vol. 14, no. 5, pp. 335–350, 2006.
- [19] G. De Jong, A. Daly, M. Pieters, S. Miller, R. Plasmeijer, and F. Hofman, "Uncertainty in traffic forecasts: literature review and new results for the netherlands," *Transportation*, vol. 34, no. 4, pp. 375–395, 2007.
- [20] T. Park and S. Lee, "A bayesian approach for estimating link travel time on urban arterial road network," in *Computational Science and Its Applications ICCSA 2004*, ser. Lecture Notes in Computer Science, A. Lagan, M. Gavrilova, V. Kumar, Y. Mun, C. Tan, and O. Gervasi, Eds. Springer Berlin Heidelberg, 2004, vol. 3043, pp. 1017–1025.
- [21] S. Sun, C. Zhang, and G. Yu, "A bayesian network approach to traffic flow forecasting," *Intelligent Transportation Systems, IEEE Transactions on*, vol. 7, no. 1, pp. 124–132, 2006.
- [22] S. Sun and C. Zhang, "The selective random subspace predictor for traffic flow forecasting," *Intelligent Transportation Systems, IEEE Transactions on*, vol. 8, no. 2, pp. 367–373, 2007.
- [23] Y. Gao, S. Sun, and D. Shi, "Network-scale traffic modeling and forecasting with graphical lasso," in *Advances in Neural Networks ISNN 2011*, ser. Lecture Notes in Computer Science, D. Liu, H. Zhang, M. Polycarpou, C. Alippi, and H. He, Eds. Springer Berlin Heidelberg, 2011, vol. 6676, pp. 151–158.

- [24] N. Djuric, V. Radosavljevic, V. Coric, and S. Vucetic, "Travel speed forecasting by means of continuous conditional random fields," *Transportation Research Record: Journal of the Transportation Research Board*, no. 2263, pp. 131–139, 2011.
- [25] J. Van Lint, "Reliable real-time framework for short-term freeway travel time prediction," *Journal of transportation engineering*, vol. 132, no. 12, pp. 921–932, 2006.
- [26] C. van Hinsbergen, J. van Lint, and H. van Zuylen, "Bayesian committee of neural networks to predict travel times with confidence intervals," *Transportation Research Part C: Emerging Technologies*, vol. 17, no. 5, pp. 498 – 509, 2009, <http://www.sciencedirect.com/science/article/pii/S0968090X09000357>.
- [27] S. Sun and X. Xu, "Variational inference for infinite mixtures of gaussian processes with applications to traffic flow prediction," *Intelligent Transportation Systems, IEEE Transactions on*, vol. 12, no. 2, pp. 466–475, 2011.
- [28] Y. Xie, K. Zhao, Y. Sun, and D. Chen, "Gaussian processes for short-term traffic volume forecasting," *Transportation Research Record: Journal of the Transportation Research Board*, no. 2165, pp. 69–78, 2010.
- [29] X. Fei, C.-C. Lu, and K. Liu, "A bayesian dynamic linear model approach for real-time short-term freeway travel time prediction," *Transportation Research Part C: Emerging Technologies*, vol. 19, no. 6, pp. 1306–1318, 2011.
- [30] B. Ghosh, B. Basu, and M. OMahony, "Bayesian time-series model for short-term traffic flow forecasting," *Journal of transportation engineering*, vol. 133, no. 3, pp. 180–189, 2007.
- [31] E. Mazloumi, G. Rose, G. Currie, and S. Moridpour, "Prediction intervals to account for uncertainties in neural network predictions: Methodology and application in bus travel time prediction," *Engineering Applications of Artificial Intelligence*, vol. 24, no. 3, pp. 534–542, 2011.
- [32] E. Mazloumi, G. Rose, G. Currie, and M. Sarvi, "An integrated framework to predict bus travel time and its variability using traffic flow data," *Journal of Intelligent Transportation Systems*, vol. 15, no. 2, pp. 75–90, 2011.
- [33] A. Khosravi, E. Mazloumi, S. Nahavandi, D. Creighton, and J. Van Lint, "Prediction intervals to account for uncertainties in travel time prediction," *Intelligent Transportation Systems, IEEE Transactions on*, vol. 12, no. 2, pp. 537–547, 2011.
- [34] —, "A genetic algorithm-based method for improving quality of travel time prediction intervals," *Transportation Research Part C: Emerging Technologies*, vol. 19, no. 6, pp. 1364–1376, 2011.
- [35] A. Khosravi, S. Nahavandi, D. Creighton, and A. F. Atiya, "Comprehensive review of neural network-based prediction intervals and new advances," *Neural Networks, IEEE Transactions on*, vol. 22, no. 9, pp. 1341–1356, 2011.
- [36] H. S. Mahmassani, "Dynamic network traffic assignment and simulation methodology for advanced system management applications," *Networks and Spatial Economics*, vol. 1, no. 3-4, pp. 267–292, 2001.
- [37] B. Park, J. Lee, D. Pampati, and B. Smith, "Online implementation of dynamit: A prototype traffic estimation and prediction program," *KSCE Journal of Civil Engineering*, vol. 12, no. 2, pp. 129–140, 2008.
- [38] C. Antoniou, *On-line calibration for Dynamic Traffic Assignment models- Theory, methods and application*. Saarbrücken, Germany, Germany: VDM Verlag, 2007.
- [39] C. Antoniou, H. N. Koutsopoulos, M. Ben-Akiva, and A. S. Chauhan, "Evaluation of diversion strategies using dynamic traffic assignment," *Transportation Planning and Technology*, vol. 34, no. 3, pp. 199–216, 2011.
- [40] S. Travis Waller and A. K. Ziliaskopoulos, "A chance-constrained based stochastic dynamic traffic assignment model: Analysis, formulation and solution algorithms," *Transportation Research Part C: Emerging Technologies*, vol. 14, no. 6, pp. 418–427, 2006.
- [41] T. Heskes, "Practical confidence and prediction intervals," in *Advances in Neural Information Processing Systems 9*. MIT press, 1997, pp. 176–182.
- [42] D. R. Cox, "Prediction intervals and empirical bayes confidence intervals," in *Perspectives in Probability and Statistics*. Academic Press, 1975, pp. 45–55.
- [43] T. Toledo, H. N. Koutsopoulos, and K. Ahmed, "Estimation of vehicle trajectories with locally weighted regression," *Transportation Research Record: Journal of the Transportation Research Board*, no. 1999, pp. 161–169, 2007.
- [44] A. Boukouvalas, R. Barillec, and D. Cornford, "Gaussian process quantile regression using expectation propagation," in *In 29th International Conference on Machine Learning (ICML-12)*. Edinburgh, Scotland, UK, 2012, 2012.
- [45] K. De Brabanter, J. De Brabanter, J. A. K. Suykens, and B. De Moor, "Approximate confidence and prediction intervals for least squares support vector regression," *Neural Networks, IEEE Transactions on*, vol. 22, no. 1, pp. 110–120, 2011.
- [46] T. Hastie and R. Tibshirani, "Generalized additive models," *Statistical science*, pp. 297–310, 1986.
- [47] R. Koenker, *quantreg: Quantile Regression*, 2013, r package version 4.98. [Online]. Available: <http://CRAN.R-project.org/package=quantreg>
- [48] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2013. [Online]. Available: <http://www.R-project.org/>
- [49] K. Yu and M. C. Jones, "Local linear quantile regression," *Journal of the American Statistical Association*, vol. 93, no. 441, pp. 228–237, 1998.
- [50] S. Geisser, *Predictive Inference: An Introduction*. CRC Press, 1993.
- [51] W. Burghout, H. N. Koutsopoulos, and I. Andreasson, "Hybrid mesoscopic-microscopic traffic simulation," *Transportation Research Record: Journal of the Transportation Research Board*, no. 1934, pp. 218–225, 2005.
- [52] H. Dette and S. Volgushev, "Non-crossing non-parametric estimates of quantile curves," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 70, no. 3, pp. 609–627, 2008.
- [53] V. Chernozhukov, I. Fernández-Val, and A. Galichon, "Quantile and probability curves without crossing," *Econometrica*, vol. 78, no. 3, pp. 1093–1125, 2010.

**Francisco C. Pereira** is Senior Research Scientist in Singapore-MIT Alliance for Research and Technology, Future Urban Mobility Integrated Research Group (FM/IRG), where he is working in real-time traffic prediction, behavior modeling, and advanced data collection technologies. In these projects, he applies his research on data mining and pattern recognition applied to unstructured contextual sources (e.g. web, news feeds, etc.) to extract information relevant to mobility. He is also a Professor in the University of Coimbra, Portugal from where he has a PhD degree, and a co-founder of the Ambient Intelligence Lab of that University. He led the development of several projects that apply these context mining principles, some with collaboration other MIT groups such as the Senseable City Lab, where he worked earlier as a postdoctoral fellow.

**Constantinos Antoniou** is Assistant Professor in the National Technical University of Athens (NTUA), Greece. He holds a Diploma in Civil Engineering from NTUA (1995), a MS in Transportation (1997) and a PhD in Transportation Systems (2004), both from MIT. His research focuses on modelling and simulation of transportation systems, Intelligent Transport Systems (ITS), calibration and optimization applications, road safety and sustainable transport systems. He has authored more than 150 scientific publications, including more than 45 papers in international, peer-reviewed journals. He is a member of several professional and scientific organizations, editorial boards (incl. Transportation Research – Part C and the International Journal of Transportation) and committees (incl. TRB committees AND20 – User Information Systems and AHB45 – Traffic Flow Theory). For more information please see <http://users.ntua.gr/antoniou>; contact him at [antoniou@central.ntua.gr](mailto:antoniou@central.ntua.gr).

**Moshe Ben-Akiva** is the Edmund K. Turner Professor of Civil and Environmental Engineering at the Massachusetts Institute of Technology (MIT), and Director of the MIT Intelligent Transportation Systems (ITS) Lab. He holds a Ph.D. degree in Transportation Systems from MIT and 4 honorary degrees and coauthored the textbook Discrete Choice Analysis.