Transportation Research Part C 96 (2018) 160–169

Contents lists available at ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

A methodology for evaluating the performance of model-based traffic prediction systems

Gabriel Gomes*, Qijian Gan, Alexandre Bayen

Institute of Transportation Studies, U.C. Berkeley, United States

ARTICLE INFO

Keywords:
Traffic simulation
Model calibration
Validation
Traffic prediction
Corridor management

ABSTRACT

Model-based traffic prediction systems (mbTPS) are a central component of the decision support and ICM (integrated corridor management) systems currently used in several large urban traffic management centers. These models are intended to generate real-time predictions of the system's response to candidate operational interventions. They must therefore be kept calibrated and trustworthy. The methodologies currently available for tracking the validity of a mbTPS have been adapted from approaches originally designed for off-line operational planning models. These approaches are insensitive to the complexity of the network and to the amount and quality of the data available. They also require significant human intervention and are therefore not suitable for real-time monitoring. This paper outlines a set of criteria for designing tests that are appropriate for the mbTPS task. It also proposes a test that meets the criteria. The test compares the predictions of the mbTPS in question to those of a model-less alternative. A t-test is used to determine whether the predictions of the mbTPS are superior to those of the model-less predictor. The approach is applied to two different systems using data from the I-210 freeway in Southern California.

1. Introduction

Short-term traffic predictions, in which the behavior of a transportation network over the next few minutes or hours is sought, are of interest to both travelers and traffic operators. There exist today a number of systems for obtaining traffic predictions. Most of them are intended to help drivers in their daily commutes. Systems such as Waze have become indispensable travel companions for the modern commuter. There is also a growing interest in traffic prediction systems to support traffic operators in their decision-making tasks. In 2006 the U.S. Department of Transportation launched the Integrated Corridor Management (ICM) initiative (Integrated corridor management, 2017) in an effort to develop new technologies to increase coordination amongst the various systems and jurisdictions that operate along a typical transportation corridor. The system design that emerged from this initiative involves a hierarchy of traffic models that work together to produce short-term forecasts of system performance under various candidate control strategies. These forecasts are used by operators to make real-time operational decisions, and must therefore be both reliable and fast.

The problem addressed in this work relates to guidelines for determining whether a particular simulation model is sufficiently well calibrated to be used either in real-time operations, or for off-line planning studies.

A 2014 Science and Policy Report by the Joint Research Centre (JRC) of the European Commission (Antoniou et al., 2014) examined the guidelines used in several countries to evaluate traffic simulation models. The report found that only about 45% of modelers polled followed guidelines of any type in evaluating their models. This low adoption rate is due in part to the fact that there

* Corresponding author.
E-mail address: gomes@path.berkeley.edu (G. Gomes).

https://doi.org/10.1016/j.trc.2018.09.004
Received 16 March 2017; Received in revised form 3 July 2018; Accepted 6 September 2018
Available online 25 September 2018

Table 1
The FHWA model assesment method.

Range	Acceptable error	Test
	Hourly link flows	
Links/times with f < 700 vph	100 vph	85%
Links/times with f ∈ [700,2700] vph	15%	85%
Links/times with f > 2700 vph	400 vph	85%
Sum of all link flows	5%	n/a
All links	GEH < 5	85%
Sum of all link flows	GEH < 4	n/a
	Travel times	
Journey time > 400 s	1 min	85%
Journey time > 400 s	15%	85%
	Visual audits	
Individual link speeds	Inspect speed-flow relationship	
Bottlenecks	Inspect queuing	

is no single accepted methodology for evaluating the veracity of traffic simulation models. In the United States, the standard method is provided by the FHWA's Traffic Analysis Toolbox (Alexiadis and Sallman, 2012; Dowling et al., 2004). This method inherits concepts originally developed in the U.K., such as the use of the GEH statistic. It consists of a series of tests that the model must pass. These tests (listed in Table 1) are of two types: error evaluation and visual audits. The error evaluation tests differ from each other in the quantities being compared (hourly link flows, sum of link flows, travel times, etc.), the error metric (absolute difference, GEH), and the thresholds (e.g. within 100 vph 85% of the time, GEH < 5, etc.). The visual audits require inspection and approval of certain diagrams by an expert.

The FHWA method is currently the only widely accepted standard for evaluating traffic simulation models in the United States. It does not, however, have general applicability for reasons listed below.

1. The test was designed for use in relatively simple freeway studies. Today, traffic simulation models are used in a wide variety of scenarios, from intersection studies, to arterial corridors, to entire cities. Tests such as the inspection of speed-flow relations and the location of bottlenecks apply only to freeways.
2. The test is inherently off-line. Real-time applications such as ICM require continuous monitoring of the quality of predictions. In this context it is not practical to require visual inspections.
3. The test is link-based, reflecting a bias toward macroscopic models. The use of the GEH statistic has been criticized for similar reasons (Antoniou et al., 2014).
4. The thresholds used in the test are fixed and insensitive to both the quality of the data and the size of the network.
5. The test applies homogeneously to the entire network. In calibrating large networks, it is usually not practical to give equal attention to all areas. Instead, modelers tend to focus their effort on important routes, such as the main arterials and strategic detour routes.

Unfortunately, this question has received little attention in the academic literature. There is a large body of work on the calibration of traffic simulation models, e.g. (Park and Schneeberger, 2003; Gomes and May, 2004). These papers usually include an error-based criterion for tracking the progress of the calibration. The task is considered complete when the error stops changing, or when it reaches a pre-established threshold (e.g. Table 1). There has also been significant work on the related topic of comparing two or more prediction methodologies, e.g. (Toledo and Koutsopoulos, 2004; Guo and Huang, 2014). Here the comparisons are usually done in terms of prediction error, and hypothesis tests are used to select a statistically significant winner.

The present work is more closely related to the first class of problems, in which a single traffic prediction system is evaluated. We address the problems identified above by proposing a set of criteria for calibration tests (Section 3), and describing a test that meets those criteria (Section 4). Sections 5 and 6 provide sample applications of the test to two different prediction systems using data from I-210 in Southern California. We begin in the next section with a generic description of the system under consideration.

2. Traffic prediction systems

We distinguish between two types of traffic prediction systems: those that involve a *mechanistic* model of the transportation network (*model-based*), and those that do not (*model-less*). There are several types of mechanistic traffic models. There are microscopic models such as Aimsun (2017) and Vissim (2017), that produce detailed trajectories of individual vehicles, mesoscopic models that produce individual travel times but not trajectories, and macroscopic models, that produce only aggregate measures of traffic such as link densities and flows. The unifying property of these model types is that they are based on traffic-specific physical principles, such as car-following equations or the hydrodynamic theory of traffic. This allows them to compute performance estimates for hypothetical scenarios. In contrast, model-less prediction systems, as defined for the purposes of this paper, are based on correlations

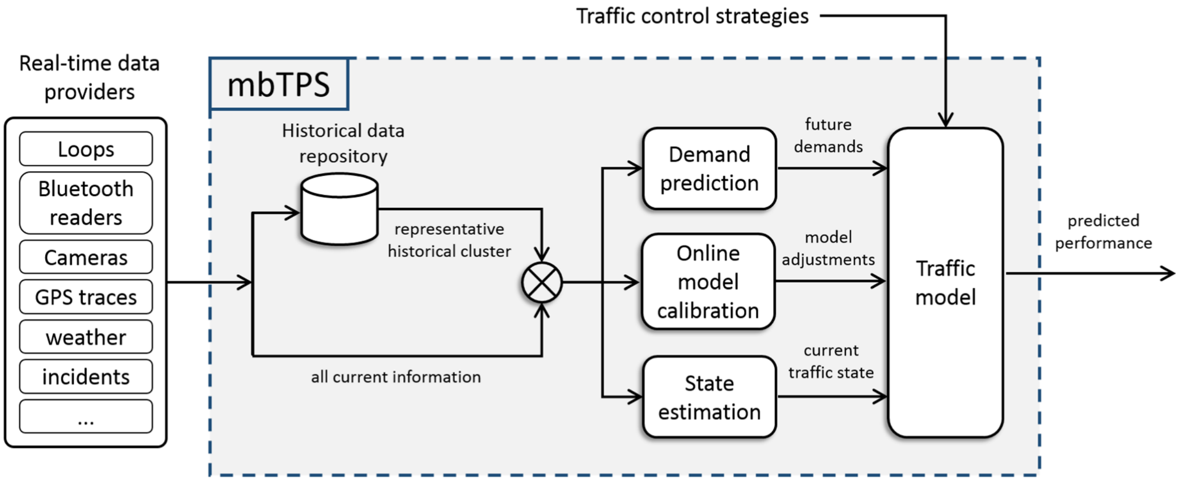

Fig. 1. A generic model-based traffic prediction system.

between the current traffic measurements and historically observed traffic states. These systems are known to provide excellent predictions of recurring traffic conditions, however they do not generate performance estimates for novel "what-if" scenarios. Such systems employ statistical rather than physical principles and are typical of commuter routing services.

Here we will describe a methodology for assessing the performance of *model-based traffic prediction systems* (mbTPS). A generic diagram for a mbTPS is shown in Fig. 1. Any particular mbTPS may or may not include all of the components depicted in the diagram, and some may include components not considered here. For example, prediction systems based on dynamic traffic assignment, such as DYNAMIT (Ben-Akiva et al., 2010), will include an additional route computation module. Our aim here is not to capture all possible mbTPS architectures, but to describe their high level operation, and to propose a test that can be applied to such systems in general.

The diagram of Fig. 1 shows the mbTPS within the dashed box. The system receives real-time traffic information from a variety of sources, including but not limited to the ones shown in the figure.

A common use of an mbTPS is to test the expected performance of a candidate traffic control strategy over some future time horizon, typically up to one hour. For example, in the event of an accident on a freeway, a traffic operator may wish to determine whether traffic diverted onto the city streets will be accommodated by the current signal plans, or if an alternative signal plan is needed. In this situation the input from the real-time data providers includes the location and severity of the incident. The mbTPS must evaluate the performance of the alternative signal plan with respect to the current plan. The traffic prediction system receives all of the current information and searches within its historical database for similar conditions seen in the past. If such a scenario is found, it will serve as a template for predicting the future evolution of the system.

The relevant historical and real-time information is provided to three modules: the demand prediction module, the state estimation module, and the online model calibration module. The purpose of these three modules is to produce the necessary input for the traffic model to create a realistic representation of the upcoming period of traffic. The *state estimator* determines the initial condition for the traffic model. That is, it computes the current arrangement of vehicles (or vehicle densities in the case of a macroscopic model) on the traffic network. This arrangement of vehicles is then used to initialize the traffic model. The *demand predictor* uses current and historical information to predict the traffic inputs for the traffic model over the prediction horizon. The implementation details of both of these modules will depend on the needs and capabilities of the traffic model. For example, if the traffic model operates with OD (origin-destination) tables, then the demand predictor will perform some form of OD estimation. If the model distinguishes between different modes of travel (vehicular, bus, truck, etc.), then the demand predictor and the state estimator should likewise provide estimates for each mode. The purpose of the *online model calibration* module is to make adjustments to the model parameters that reflect the current environmental conditions. For example, precipitation may be represented in the model by decreasing the expected speed of vehicles. Traffic incidents may be represented by adjusting the relevant model parameters according to the location and severity of the incident. As with the other two modules, the details of the online calibration module will depend on the specifics of the model.

The *traffic model* is initialized with the outputs of the demand prediction, calibration, and state estimation modules. These outputs may be deterministic or probabilistic. For example, the estimated state may be specified as the exact number of vehicles in each link, or as probability distributions representing the range of possible values given the measurements and their inherent uncertainty. Similarly the predicted demands and calibrated model parameters may be specified as numbers or as distributions.

The traffic model then runs some number of times to produce a prediction of the performance of the system under the candidate control strategy. The performance metrics produced by the system will depend on the needs of the operator and on the capabilities of the model. In the design of the mbTPS assessment test we assume only that the performance metric can be computed from the data supplied by the real-time data providers. That is, the ability of the system to predict the upcoming period of traffic in terms of some performance metric must be verifiable from data collected after the period has elapsed.

A list of the existing traffic prediction systems can be found in (Wikipedia, 2017). The systems that were created as part of the USDOT ICM initiative for the U.S. 75 corridor in Dallas and Interstate 15 in San Diego are documented in (USDOT, 2017).

3. Design criteria for mbTPS tests

Traffic prediction systems are only useful if their predictions can be trusted. This means that the difference between the predicted and actual performance of the traffic network – the prediction error – should be expected to be small. Hence it is important to evaluate these errors by means of a test. The main considerations when designing such a test are (a) how to calculate the error and (b) how to define "small". The following is a list of criteria which we consider important for any mbTPS assessment test.

1. *It should be applicable to any network*. It should apply equally to freeways, arterials, or large network models.
2. *It should be system agnostic*. The test should not make any assumptions regarding the internal components of the mbTPS beyond the high-level architecture described in the previous section. For example, the test should apply equally to demand/split based models and to OD/route based models. It should also apply equally to macroscopic, mesoscopic, and microscopic traffic models.
3. *It should adapt with data quality*. The quality and availability of real-time data vary from day to day and from one facility to another. Hence, the quality of the predictions based on that data can also be expected to vary. A test that is insensitive to such variations will be incapable of distinguishing between errors caused by bad data and errors caused by bad modeling. On the other hand, a test that adapts itself to the quality of the data will help the operator to determine the sources of error, and thus to allocate resources toward either sensor improvements or model calibration.
4. *It should directly evaluate the ability of the mbTPS to predict traffic performance metrics*. It is not sufficient to test each of the components of the system in isolation. The test must ultimately evaluate the ability of the mbTPS to perform its main task, which is to predict the behavior of the traffic system under previously unobserved conditions.
5. *Its conclusions should be statistically sound*. The statements produced by the test (e.g. "system A passes the test"), should be accompanied by a measure of their statistical significance.
6. *It should be simple and inexpensive to perform*. Ideally the test should run in real time alongside the mbTPS, and produce a continuous check on its performance. The test should not require significant human intervention.

It should be noted that the FHWA method complies only with the criterion #2.

1. It applies only to freeway networks.
2. Its threshold do not adapt with the quality of the data.
3. It evaluates flows and travel times on all links, whereas the operator may be more interested in other performance metrics, such as the total delay or the queues along a given route.
4. The test is based on a single sample day. There is no accounting for day-to-day or week-to-week variations.
5. It requires a visual audit by an analyst.

4. A test based on model-less prediction

Two important shortcomings of the FHWA standard should be emphasized: (a) its thresholds are fixed, and (b) it lacks statistical significance. To address the first problem we introduce an alternative *model-less* predictor: the *mlTPS*. This mlTPS is given the same prediction task as the mbTPS, and its resulting prediction errors used in place of fixed thresholds. The second problem, the lack of statistical significance, is addressed with a hypothesis test.

The proposed framework is illustrated in Fig. 2. The system being evaluated is the shaded block. The test harness receives some amount of traffic information from its data providers, shown on the left side of the diagram. This data is evaluated by a data quality assessment module, which appends a health score to each of the data packets. There is a large flexibility regarding the specificity of

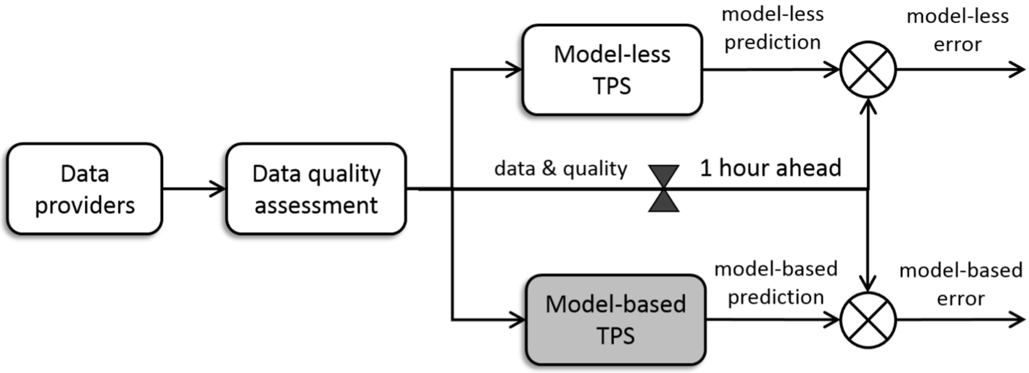

Fig. 2. Model-based TPS testing framework.

the data health assessment. For example, a single score may be attached to an entire day of measurements from each detector station, as in the PeMS database (PeMS, 2017). Or data health may be evaluated on an hourly or even continuous basis. The stream of data along with its score is provided to the mbTPS being evaluated and also to the model-less TPS. Thus, both systems receive exactly the same real-time information.

The proposed test can be informally stated as follows,

> A mbTPS passes the test if it can be said with certainty γ that the prediction errors it produces are smaller than those produced by a mlTPS by an amount λ.

4.1. The model-less TPS

The mlTPS is considered an inexpensive means of producing predictions of performance. It is inexpensive relative to the model-based system because it avoids the time-consuming and labor-intensive process of building and calibrating a traffic model. The proposed test asserts that the model-building effort should produce a system with an improved ability to predict traffic. Hence the predictive capabilities of the model-less system establish a bound for the performance of the model-based TPS. In designing the reference model-less TPS, we propose the following criteria,

1. It should not include any *mechanistic* model of the traffic system.
2. It should access the same data and data health information as the mbTPS.
3. It should not require any details of the traffic network or of the traffic control system beyond what is used for the calculation of performance metrics.
4. It may have access to the historical database and traffic state clustering system.

These criteria are designed to exclude systems that require a large effort to set up and configure and are therefore not aligned with the concept of an "inexpensive alternative". However, they do not exclude systems based on fairly sophisticated techniques of data mining and statistical learning.

We illustrate the approach using the simple model-less TPS described by Eq. (1). This formula assumes that the data health score has been provided as a scalar $r_i(t) \in [0, 1]$ for each data source (e.g. loop detector) i and time instant t. The value $r_i(t)$ ranges from 0 for useless data, to 1 for data that is completely reliable. This model-less predictor also has access to a representative profile for each data source obtained from the historical database. The prediction of any measured quantity $x_i(t)$ is then given by,

$$\hat{x}_i(t+h|t) = r_i(t)x_i^{rt}(t) + (1-r_i(t))x_i^{hist}(t+h) \tag{1}$$

Here $\hat{x}_i(t+h|t)$ represents the prediction of quantity x_i at time $t+h$ using information up to time t. The symbol "x" may stand for flow, occupancy, speed, or any other measured quantity. $x_i^{rt}(t)$ is the real-time measurement of $x_i(t)$, and $x_i^{hist}(t+h)$ is the representative historical value of x_i at time $t+h$. As mentioned above, the computation of $x_i^{hist}(t+h)$ may engage a statistical model. The equation selects either the historical or the real time estimate in the extreme cases of $r_i(t)=0$ and $r_i(t)=1$, and computes a linear combination of the two in intermediate cases.

The complete model-less TPS consists of evaluating Eq. (1) for every data source and applying the performance metric calculation to the result. It is assumed, as was mentioned in Section 2, that the performance metrics can be computed from the measurements. Thus, a performance calculator is available that maps traffic measurements to performance metrics. This component is not considered as part of the model-less TPS, and is therefore allowed to include information about the traffic network, including road geometries and speed limits.

To give an example, assume that the performance calculator provides the following formula for Vehicle Miles Traveled (VMT) incurred over a period $[t, t+h]$,

$$\text{VMT} = \sum_i L_i \int_t^{t+h} f_i(\tau)d\tau \tag{2}$$

Here the summation is made over all sensors i. L_i is the length of road represented by sensor i, and f_i is flow. The performance calculator contains the L_i's and evaluates the function on a set of flow profiles $\{f_i\}$. This function can also be evaluated on a set of *predicted* flow profiles. For example, profiles computed with Eq. (1), which leads to the following prediction of VMT,

$$\widehat{\text{VMT}}\left(\left[t, t+h\right]\middle|t\right) = \sum_i L_i\left(h\, r_i(t)f_i^{rt}(t) + \left(1-r_i(t)\right)\int_t^{t+h} f_i^{hist}(\tau)d\tau\right) \tag{3}$$

4.2. Hypothesis testing

The informal test statement provided in the previous section can be made more precise by casting it in terms of a hypothesis test (Triola, 2006). The statistical unit for the test is a combination of the road network with a full day of traffic data and external factors such as accidents and weather. For example, interstate 210 East on March 20th, 2016 constitutes a single statistical unit. The two traffic prediction systems – the mbTPS and the mlTPS – are regarded as separate interventions performed on a common set of statistical units (i.e. over several days). The prediction errors produced by the two traffic prediction systems are the test outcomes.

This setup is analogous to that of a group of subjects being administered two separate medical treatments, with the goal being to determine whether one treatment is more effective than the other. A well-known strategy for such experiments is to use a paired sample, one-tailed t-test. A t-test is used whenever the outcomes of the two interventions can be assumed to be normally (or almost normally) distributed. Paired tests are used when the samples can be organized into correlated pairs, such as in the present case, in which the two systems are applied to a common set of statistical units. The one-tailed version of the test is used when the alternative hypothesis states that one of the interventions is superior to the other.

The test is to be performed on a set of traffic data covering a sufficiently large period of time. It may also be considered important to include a sufficiently rich variety of traffic scenarios, including light, moderate, and severe congestion, with and without accidents, etc. For example, one may consider running the test over an entire month of traffic data. The amount of data included in the test will influence its ability to reject the null hypothesis for a given level of significance.

We consider an example in which 30 days of traffic data are used. The two traffic prediction systems (mbTPS and model-less) run every 30 min for the 30 days, resulting in a total of 1,440 (48 x 30) predictions for each one. We use $e_{d,k}^{mb}$and $e_{d,k}^{ml}$ ($d = 1...30$, $k = 1...48$) to denote the error from the k'th prediction period of the d'th day for the model-based and model-less predictors respectively. e_d^{mb} and e_d^{ml} denote the average prediction error for the d'th day, computed by taking the mean of the 48 predictions for that day. e_d^{mb} and e_d^{mb} represent paired samples for d'th statistical unit.

The proposed null hypothesis:

$$H_0: \qquad E(e_d^{mb}) = (1+\lambda)E(e_d^{ml}) \tag{4}$$

asserts that the means of the errors obtained with the model-based predictor are equal to $(1+\lambda)$ times those of the model-less system. With λ set to zero, H_0 asserts that expected errors of the two systems are the same. H_0 is weighed against the alternative hypothesis,

$$H_1: \qquad E(e_d^{mb}) < (1+\lambda)E(e_d^{ml}) \tag{5}$$

that the expected prediction errors with the model based system are better than the model-less system, inflated by λ percent. The parameter λ controls the difficulty of the test: tests with positive λ are easier to pass than tests with negative λ. This is illustrated in Fig. 3. This problem is cast as a paired test using the statistic $y_d = e_d^{mb} - (1+\lambda)e_d^{ml}$. Then the null and alternative hypotheses become H_0: $E(y) = 0$ and H_1: $E(y) < 0$.

The t statistic for this test is calculated with,

$$t_y = \frac{\overline{y_d}}{s_{y_d}/\sqrt{n}} \tag{6}$$

Here $\overline{y_d}$ and s_{y_d} are respectively the sample mean and variance of $\{y_d\}$ and n is the sample size (30 in this case). The sample mean is assumed to be distributed according to a t distribution with $n-1$ degrees of freedom. The null hypothesis is rejected in favor of the alternative hypothesis in the case that the likelihood of obtaining an outcome as or more extreme than t_y (the p-value) is smaller than a pre-established threshold (the statistical significance α, usually 0.05 or 0.01).

We can additionally compute the γ confidence interval for the mean of the true distribution of y. In the case of a one-sided t-test, this interval is of the form $[\delta, \infty)$. Its interpretation is that, with confidence level γ, the true mean of y is greater than δ. The computation of δ is as follows.

$$\delta = \overline{y_d} - t_c\frac{s_{y_d}}{\sqrt{n}} \tag{7}$$

where t_c is the value for which γ percent of the zero-mean t distribution with $n-1$ degrees of freedom is contained between $-t_c$ and t_c.

The model-based TPS is considered to have passed the test when the null hypothesis can be rejected in favor of the alternative hypothesis with a p-value of at most $\alpha = 1-\gamma$ (e.g. 5%). In this case, the value $\overline{y_d} - \delta$ represents a conservative estimate of the improvement afforded by the mbTPS.

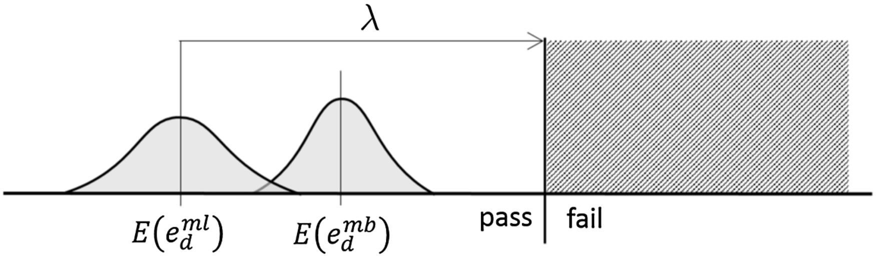

Fig. 3. Paired sample, one-tailed t-test.

4.3. Comments on the method

This test framework complies with each of the requirements listed in Section 3.

1. *It is applicable to any network*. All that is needed to run the test is a suitable model-less predictor. Because this predictor does not contain a traffic model, it is independent of the topology of the network.
2. *It is system agnostic*. The evaluation framework treats the mbTPS as a black box. That is, it provides inputs and computes errors based on the outputs. Hence it makes no assumptions regarding the internal structure of the mbTPS.
3. *It adapts with data quality*. The quality of the mbTPS is evaluated relative to the model-less TPS, and not relative to a fixed threshold. As the quality of the data decays, so will the quality of the predictions of the model-less TPS. This allows the operator to distinguish between data-induced and model-induced prediction errors.
4. *It directly evaluates the ability of the mbTPS to predict traffic performance metrics*. In the presented methodology, errors can be evaluated on any quantity that can be computed from measurements. This will necessarily include all traffic performance measures of interest.
5. *Its conclusions are statistically sound*. Although the approach was illustrated with a t-test, other forms of hypothesis test apply to more complex distributions. Hence, an appropriate test can be designed for different measures of performance. Also, the desired level of significance is set beforehand by the analyst. This parameter will determine the minimum number of runs needed to pass the test (another concern reported in (Antoniou et al., 2014)).
6. *It is simple and inexpensive to perform*. The methodology is fully numerical and does not require any subjective evaluation by an expert. Thus it is suitable for real-time implementation.

5. Application to I-210 Eastbound

We demonstrate the methodology with two candidate traffic prediction systems for the I-210 Eastbound freeway. The two systems, referred to as mbTPS-A and mbTPS-B, are instances of the general configuration shown in Fig. 1, with components described in Table 2.

It can be seen from the table that the two systems differ only in the algorithms used for demand and split ratio prediction. These modules produce the 1-h estimates of future demands and split ratios that are provided to the traffic simulator in order to make predictions of traffic performance. mbTPS-A uses the “Hold or Historical” algorithm, meaning that, similarly to the model-less system, it holds the current flow measurement if the sensor is healthy, and uses a historical profile if it is not healthy. This approach works well during periods of relatively steady traffic, but fails during transition periods preceding and following the peak periods. mbTPS-B uses a more sophisticated time-series prediction algorithm described in (Wu et al., 2014a,b).

Both mbTPS-A and mbTPS-B use simple linear interpolation to estimate densities on the freeway. This means that the instantaneous density profile on the freeway is interpolated from point measurements at mainline loop detector stations. The two systems use identical traffic simulation models: the Cell Transmission Model (CTM) as implemented in the BeATS simulator (OTM, 2018).

The I-210 East freeway network consists of 92 VDSs (vehicle detector stations), including 36 mainline stations, 24 off-ramps, and 29 on-ramps, and 3 freeway connector stations. Not all of these stations were in good working conditions during the analysis. For each day we used the PeMS loop health diagnostic to determine the set of working sensors. This set changed from day to day, but on average there were about 79 good stations per day: 36 mainline, 16 offramp, 25 onramp, and 2 freeway connectors.

For the analysis we selected a set of 35 “good days”. This selection was done manually by inspecting the health of the detectors, as well as the congestion patterns. A variety of congestion patterns and days of week were selected, including weekend days, and days with obvious incidents. This was done in order to test the systems over a range of scenarios.

The data provided to both traffic prediction systems was the 5-min readings from all of the healthy stations, as well as historical profiles. These historical profiles were computed by taking the average of the 35 days of data for each day of the week.

Table 2
Configurations for two traffic prediction systems.

Component	mbTPS-A	mbTPS-B
Historical data clustering	Day-of-week	Day-of-week
State estimator	Linear interpolation	Linear interpolation
Demand predictor	Hold or Historical	ARMAX
Split ratio predictor	Hold or Historical	ARMAX
Traffic model	CTM	CTM
Online calibration	None	None

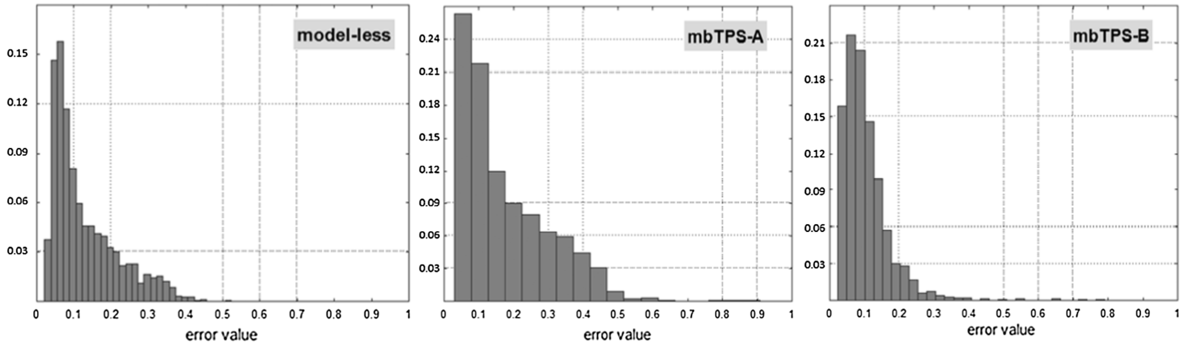

Fig. 4. Histogram of prediction errors computed with Eq. (8).

6. Results

6.1. Experiment with full sensing

The two systems were used to generate 1-h predictions every 30 min for the 35 days. The error for each of the 1-h prediction period was calculated as follows,

$$e_{d,k} = \frac{1}{n_S}\sum_{i=1}^{n_S} \frac{|f^{pred}_{d,k,i} - f^{meas}_{d,k,i}|}{f^{meas}_{k,i}} \tag{8}$$

That is, as the mean absolute percentage error of the predicted flows $f^{pred}_{d,k,i}$. Here d is a day index ($d = 1...35$), k is the time of day ($k = 1...48$), i is the sensor index, and n_S is the number of healthy sensors. $f^{pred}_{d,k,i}$ and $f^{meas}_{d,k,i}$ are respectively the predicted and measured average flow for sensor i over prediction period k on day d. The 48 prediction errors corresponding to each day were then averaged,

$$e_d = \frac{1}{48}\sum_{k=1}^{48} e_{d,k} \tag{9}$$

These prediction errors were calculated for each of the 3 traffic prediction systems. Histograms for $e_{d,k}$ are shown in Figs. 4 and for e_d in Fig. 5. Fig. 4 shows that the individual predictions errors resemble exponential distributions.

One of the assumptions of the t-test is that the underlying distribution should be normal. This is in order that the sample means follow a t distribution with $n-1$ degrees of freedom. Although this assumption is violated in our case (the distributions are approximately exponential), the large sample size (35) generates a sample mean that is approximately normal. This is verified in Fig. 6, which shows a synthetically produced histogram of means of samples of size 35 from an exponential distribution. The histogram is overlaid with a normal distribution for comparison. Although the histogram is slightly skewed to the left, it is clear that its shape is sufficiently normal to justify the use of a t-test.

The results of the two paired t-tests are provided in Table 3. The parameters used in these tests were $\gamma = 0.95$ ($\alpha = 0.05$) and $\lambda = 0$ (H_1: "the mbTPS is at least as good as the model-less TPS with 95% confidence"). The first two rows provide the basic statistics on the data: the sample means for mbTPS-A and mbTPS-B are 17.6% and 10.5% respectively. By comparison, the sample mean for the model-less TPS is 12.8%.

The next two rows provide the corresponding t-statistic and p-value. With a 95% confidence, these p-values imply a rejection of the null hypothesis for mbTPS-B, but not for mbTPS-A. Hence the confidence interval for mbTPS-A is not meaningful and is therefore not reported.

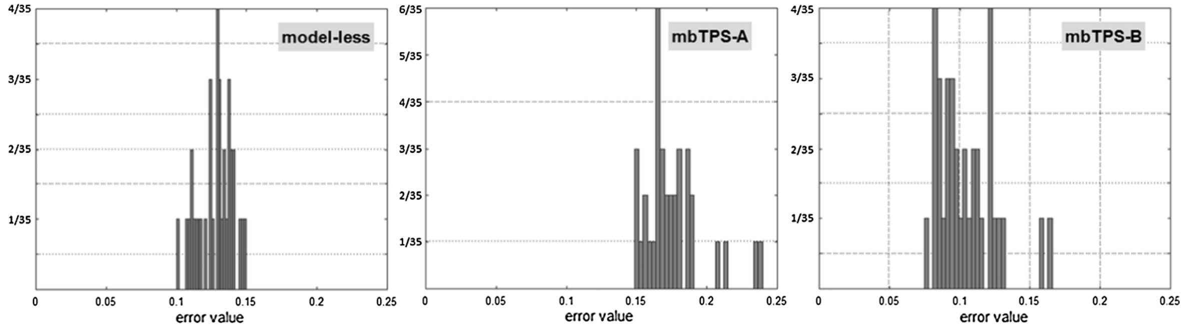

Fig. 5. Histogram of the sample, Eq. (9).

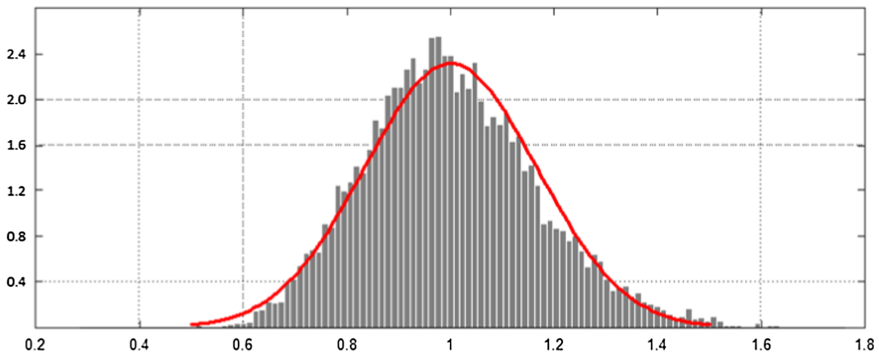

Fig. 6. Histogram of the sample mean of an exponential distribution compared to a normal distribution (sample size = 35).

Table 3
Results of the hypothesis tests.

Component	mbTPS-A	mbTPS-B
Sample mean	17.6%	10.5%
Sample standard deviation	2.1%	2.1%
t-statistic	−14.71	6.03
Rejected H0?	No	Yes
Confidence interval	–	[1.7%, ∞)

6.2. Experiment with degraded sensing

Next we demonstrate the adaptability of the test to the quality and amount of available data. Whereas in the previous section all of the data was provided to both of the prediction systems, here we mimic sensor failures by overwriting the health scores of a randomly selected set of sensors. Three degraded sensing scenarios were tested.

1. Removing 15 mainline stations and no ramp stations,
2. Removing 17 ramp stations and no mainline stations,
3. Removing 28 mainline stations and 34 ramp stations.

Given that there are typically around 30 good mainline stations and 34 good ramp stations on any given day on I-210E, these experiments respectively correspond to scenarios of 50% mainline availability, 50% ramp availability, and almost no data. Again, the null and alternative hypotheses being considered are:

H_0: $E(y_d) = 0$
versus
H_1: $E(y_d) < 0$

where $y_d = E(e_d^{mb}) = (1 + \lambda)E(e_d^{ml})$ is the pair-wise difference between the model-based and model-less prediction errors. λ establishes the pass/fail threshold, as was illustrated in Fig. 3. Observe that the test cannot be passed if λ is set to -1, since this would require $E(e^{mb}) < 0$ and the error values are always positive. On the other hand, the test can certainly be passed for some sufficiently large value of λ. Hence there is a threshold $\overline{\lambda}$ which is the minimum value of λ for which the test is passed by a given model-based system. This value gives some idea of the margin by which the system either beats or is beaten by the model-less predictor.

The scenarios were tested by randomly removing a set of stations from the set of good stations. In each case, the experiment was

Table 4
Test results with degraded sensor environments.

	Scenario	1.	2.	3.
	# ML removed	15	0	28
	# RP removed	0	17	34
	$E(e^{ml})$	13.1%	12.9%	13.2%
mbTPS-A	$E(e_d^{mb})$	15.2%	18.0%	16.4%
	$\overline{\lambda}$	0.22	0.53	0.35
mbTPS-B	$E(e_d^{mb})$	9.6%	10.2%	10.2%
	$\overline{\lambda}$	−0.16	−0.07	−0.12

repeated 20 times, providing samples of size 20 for the hypothesis test. Results from Table 4 indicate that the model based system with the zero-order hold flow predictor (mbTPS-A) cannot improve upon the model-less predictor for any level of sensor availability. On the other hand, the system with the ARMAX predictor (mbTPS-B) outperforms the model-less predictor in every case. It should be noted that these results were obtained from a very small number of trials, when compared to the large number of possible combinations. For example, the sample size of 20 for the first scenario is vanishingly small when compared to the over 150 million ways of removing 15 mainline sensors from a pool of 30. Nevertheless, the conclusion that the ARMAX-based predictor is superior to the ZOH-based predictor is consistent and significant.

7. Conclusion

This paper has described a methodology for testing the performance of a traffic prediction system. The design of the methodology was based on a set of criteria, listed in Section 3. These critera can be generalized beyond traffic prediction to cover a wider range of applications. In general, we may consider the design of a methodology ϕ for evaluating a system A, which takes measurements x as input, and produces an estimate or a prediction of some other quantity of interest y. In this context, the list of criteria can be stated as follows.

1. ϕ should apply over entire domain of A.
2. ϕ should make as few assumptions as possible about the internal structure of A.
3. ϕ should account for the quantity and quality of the input data x, and become less stringent as x degrades.
4. The conclusions of ϕ should be based primarily on y and not on any ancillary outputs of A.
5. The conclusions of ϕ should be paired with a statistical notion of their certainty.

We have left out the last criterion because it applies narrowly to real-time systems. The methodology we have proposed consists of using the outputs of a simple alternative to A, which we will denote A_o. The domain of A_o should include the domain of A (item 1). Then the outputs of A_o can be used as a reference with which to evaluate the outputs of A. A hypothesis test is then formulated to determine whether A is significantly better than A_o at computing outputs y from inputs x. Notice that this methodology satisfies all of our stated criteria.

To illustrate the generalization, consider the evaluation of an algorithm for the state estimation block of Fig. 2; a Kalman filter, for example. The methodology requires that we supply a simple alternative algorithm; for example, we might consider interpolating the traffic density along the road between measurements. We could then proceed by estimating a single measurement from all other measurements, using both linear interpolation and the Kalman filter. A hypothesis test would then determine, with explicit statistical confidence, whether the Kalman filter is better than linear interpolation by some given margin.

Acknowledgement

This work was funded by Caltrans through the U.C. Berkeley Connected Corridors Program.

References

Aimsun, https://www.aimsun.com/ (viewed December, 2017).

Alexiadis, V., Sallman, D., 2012. “Traffic Analysis Toolbox Volume XIII: Integrated Corridor Management Analysis, Modeling, and Simulation Guide”. Tech. Rep., U.S. Department of Transportation (May, 2012).

Antoniou, C., Barcelo, J., Brackstone, M., Celikoglu, H., Ciuffo, B., Punzo, V., Sykes, P., Toledo, T., Vortisch, P., Wagner, P., 2014.“Traffic Simulation: Case for guidelines”. Tech. Rep. Joint Research Centre of the European Commission; 2014.

Ben-Akiva, M., Koutsopoulos, H.N., Antoniou, C., Balakrishna, R., 2010. Traffic simulation with DynaMIT. In: Fundamentals of Traffic Simulation, Springer.

Dowling, R., Skabardonis, A., Alexiadis, V., 2004. “Traffic Analysis Toolbox Volume III: Guidelines for Applying Traffic Microsimulation Modeling Software”. Tech. Rep.. U.S. Department of Transportation (June 2004).

Gomes, G., May, A., 2004. "Congested Freeway Microsimulation Model Using VISSIM". Transport. Res. Rec.: J. Transport. Res. Board 1876, 71–81.

Guo, J., Huang, W., Williams, B., 2014. Adaptive Kalman filter approach for stochastic short-term traffic flow rate prediction and uncertainty quantification. Transport. Res. Part C: Emerg. Technol. 43 (Part 1), 50–64.

Open Traffic Models (OTM). https://github.com/ggomes/otm-sim (viewed August 2018).

Park, B., Schneeberger, J.D., 2003. Microscopic simulation model calibration and validation case study of VISSIM simulation model for a coordinated actuated signal system. Transport. Res. Record 1856.

PeMS website. http://pems.dot.ca.gov (viewed December, 2017).

Toledo, T., Koutsopoulos, H., 2004. Statistical validation of traffic simulation models. Transport. Res. Record 1876.

Triola, M.F., 2006. Elementary Statistics. 10th Edition. Pearson/Addison-Wesley.

USDOT. http://www.its.dot.gov/research_archives/icms (viewed December, 2017).

Vissim, http://vision-traffic.ptvgroup.com/en-us/products/ptv-vissim/ (viewed December, 2017).

Wikipedia. Traffic estimation and prediction system, https://en.wikipedia.org/wiki/Traffic_estimation_and_prediction_system (viewed December, 2017).

Wu, C.J., Schreiter, T., Horowitz, R., Gomes, G., 2014a. Fast boundary flow prediction for traffic flow models using optimal ARMAX based predictors. Transport. Res. Rec.: J. Transport. Res. Board 2421, 125–132.

Wu, C.J., Schreiter, T., Horowitz, R., 2014b. Multiple-clustering ARMAX-based predictor and its application to freeway traffic flow prediction. In: American Control Conference, pp. 4397–4403.