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

**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