Transportation Research Part A 64 (2014) 65–91

Contents lists available at ScienceDirect

Transportation Research Part A

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

Border crossing delay prediction using transient multi-server queueing models

Lei Lin, Qian Wang, Adel W. Sadek *

Department of Civil, Structural and Environmental Engineering, University at Buffalo, The State University of New York, Buffalo, NY 14260, USA

ARTICLE INFO

Article history:
Received 30 November 2012
Received in revised form 3 September 2013
Accepted 24 March 2014
Available online 18 April 2014

Keywords:
Border-crossing delay prediction
Multi-server queueing models
Batch markovian arrival process
Phase type
Transient solution
Traffic simulation

ABSTRACT

As a result of the continued increase in travel demand coupled with the need for tighter security and inspection procedures after September 11, border crossing delay has recently become a critical issue with tremendous economic and social costs. The current paper develops multi-server queuing models to estimate border crossing delay in support of a predictive traveler information system for the crossings. Two classes of multi-server models are considered: (1) models with exponential inter-arrival times and Erlang service times; and (2) a more generic model with a Batch Markovian Arrival Process (BMAP) and phase types (PH) services. As a case study, the models are developed based on real-time traffic volume and inspection time data collected at one of the major US–Canada border crossings, the Peace Bridge, and their transient solution is obtained using heuristic methods. For validation, the queueing models' estimates are compared to the results from a detailed microscopic traffic simulation model of the Peace Bridge border crossing. The comparison shows that the transient queueing model, along its heuristic solution algorithm, is capable of predicting border crossing delay. Finally, a set of sensitivity analysis tests are conducted, and the developed models are incorporated within an optimization framework to help inform border crossing management strategies.

1. Introduction

In recent years, as a result of the continued increase in travel demand across the border coupled with the need for tighter security and inspection procedures after September 11, border crossing delay has become a critical problem with tremendous economic and social costs. A study conducted for the US Department of Transportation back in 2003 indicated that border crossing delays cost the US and Canada more than \$13.2 billion every year (Taylor et al., 2003), and a report by the Ontario Chamber of Commerce (OCC) warned that if not adequately addressed, delay at the Detroit-Windsor border crossings may result in over 70,000 jobs lost by 2030 (OCC, 2004). More recently, in a press release in 2008, US former Transportation Secretary, Mary E. Peters, highlighted the border crossing delay problem by stating that such delays had cost businesses on the Canadian and the US sides as many as 14 billion dollars in 2007 (USDOT, Office of Public Affairs, 2008). Furthermore, a study by the Departments of Economics at the University of Waterloo and Wilfrid Laurier University in Canada concluded that the figures estimated by previous studies for the cost of the US–Canadian border crossing problem, while staggering, may have actually underestimated the true cost (Nguyen and Wigle, 2011).

* Corresponding author. Tel.: +1 716 645 4367; fax: +1 716 645 3733.
E-mail address: asadek@buffalo.edu (A.W. Sadek).

http://dx.doi.org/10.1016/j.tra.2014.03.013
0965-8564/

To alleviate the problem to some extent, transport authorities in the US and Canada presently provide *current* or *instantaneous* border-crossing delay information to the public, based on real-time traffic information collected at the border. The *current* border crossing times are communicated to the public via several information dissemination avenues, such as websites, on-road dynamic message signs and a traveler information phone system. While *current* border-crossing delay information is quite valuable in itself, it may be quite different from the travel times that the drivers would actually experience by the time they arrive at the border (i.e. *experienced* travel times). This is especially the case if there is a significant lag between the time when a traveler receives (or needs to act upon) the information and the time when she/he arrives at the border. This difference decreases the value of delay information in guiding traffic and mitigating congestion.

To address this, the current paper develops *predictive* models of border delay which can be used to provide drivers with estimates of the delay they are likely to *experience* when they arrive at the border. We view the problem of providing predictive information about border crossing delay as consisting of two sub-problems that need to be solved sequentially. The first is the short-term traffic volume forecasting problem, which is concerned with providing short-term predictions of the likely traffic volume at the border crossing for the next time period ranging from 30 min to a few hours. With this information, the second problem involves formulating and solving the appropriate queueing model in order to estimate the expected delay as a function of the predicted arrival volume from the solution of the first problem, the service rate and the number of customs inspection booths open.

Regarding the short-term traffic volume forecasting problem, extensive studies have been done for both freeway facilities (e.g. Smith and Demetsky, 1997; Park et al., 1998; Williams and Hoel, 2003; Smith et al., 2002; Huang and Sadek, 2009) and arterial streets (e.g. Lin et al., 2004; Liu et al., 2006). Moreover, for the special case of short-term predicting of border crossing traffic, the authors in a previous paper have proposed a multi-model combined forecast method which was shown to yield exceptionally good results (Lin et al., 2012).

Given this, the focus of this paper is on the second problem involving the formulation and solution of the appropriate queueing model for estimating border delay. Specifically, in this study, we formulate two classes of multi-server queueing models. The first looks at the special case of Erlang service times and exponential inter-arrival times (i.e.an $M/E_k/n$ queueing model), and the second looks at the more generic case of a Batch Markovian Arrival Process (BMAP) and phase type (PH) services (i.e., *BMAP/PH/n* queueing model). The models are developed based on data collected from the Peace Bridge, one of the four major border crossings connecting Western New York, US to Southern Ontario, Canada, and their transient solutions are derived using heuristic approaches. For validation, the transient queuing model solutions are compared to those estimated from a well-calibrated microscopic traffic simulation model built for the Peace Bridge border crossing using VISSIM (PTV, 2010). Finally, a set of sensitivity analysis tests are conducted and the developed models are incorporated within an optimization framework to help inform border crossing management strategies.

The paper is organized as following. Section 2 provides background information about queueing models and their use for online travel time prediction. Section 3 introduces the Peace Bridge border crossing case study and describes the data collected. Section 4 discusses the details of the methodology. Specifically, that Section 1 describes the $M/E_k/n$ queueing model formulation and its transient solution algorithm, followed by the *BMAP/PH/n* formulation along with its solution algorithm. It also describes the validation procedure which involved comparing the models' results to the VISSIM simulation model's results. Section 5 presents the validation and sensitivity analysis results. That section also includes an example on how the queueing models may be applied to develop ‘‘optimal'' border management strategies. Finally, the main conclusions are summarized.

2. Literature review

Queueing models have been used extensively to solve problems related to manufacturing processes, transportation systems, product distribution systems, call centers, among other applications (Gontijo et al., 2011). Queueing models can be categorized in a number of different ways. One categorization divides them into stationary queueing models and transient queueing models as explained below.

Let $N(t)$ denote the number of customers (i.e. vehicles) in the queueing system at time t measured from a fixed initial time moment $t = 0$, and let $p_n(t)$ denote the probability that $N(t) = n$ at time t. Because it is usually difficult to find the time-dependent solution $p_n(t)$ *analytically*, many applications in practice resort to consider only the steady state behavior of the queueing system after being in operation for a sufficiently long time. In that case, one is interested in the limiting behavior of $p_n = \lim_{t\to\infty} p_n(t)$, $n = 0, 1, 2, \ldots$ (Medhi, 2003). These queueing models that study the limiting probability of $p_n(t)$ are called stationary queueing models.

Stationary queueing models are usually used to derive some useful performance measures such as the average waiting time, which in turn is often adopted as an objective function within an optimization framework to improve the efficiency of a system. For example, the study by Kim (2009) built a non-linear integer programming model to study the toll plaza optimization problem. In his model, the cost of the waiting time of the vehicles, as determined from the steady state solution of the queueing model, was minimized. Another example is the study by Ausín et al. (2007) which tried to minimize the steady-state expected total waiting time cost by optimizing the number of servers based on real data from a bank. Moreover, Zhang et al. (2011) proposed a two-stage queueing model to balance security and customer service goals for a border crossing system. Besides waiting time, some researchers tried to estimate additional measures of queueing systems, such as traffic