Submitted to *Transportation Research Part A: Policy and Practice*
manuscript TRA-D-15-00506.R1

A Data-Driven Approach to Manpower Planning at U.S.-Canada Border Crossings

Mengqiao Yu*, Yichuan Ding†, Robin Lindsey†, Cong Shi‡

* Transportation Engineering, University of California, Berkeley, CA 94720, USA mengqiao.yu@berkeley.edu

† Sauder School of Business, University of British Columbia, Vancouver, V6T 1Z2, British Columbia, Canada {daniel.ding, robin.lindsey}@sauder.ubc.ca

‡ Industrial and Operations Engineering, University of Michigan, Ann Arbor, MI 48109, USA shicong@umich.edu

We investigate the staffing problem at Peace Arch, one of the major U.S.-Canada border crossings, with the goal of reducing time delay without compromising the effectiveness of security screening. Our data analytics show how the arrival rates of vehicles vary by time of day and day of week, and that the service rate per booth varies considerably by the time of day and the number of active booths. We propose a time-varying queueing model to capture these dynamics and use empirical data to estimate the model parameters using a multiple linear regression. We then formulate the staffing task as an integer programming problem and derive a near-optimal workforce schedule. Simulations reveal that our proposed workforce policy improves on the existing schedule by about 18% in terms of average delay without increasing the total work hours of the border staff.

Key words: data-driven, workforce policy, queueing, empirical data analysis
History: Received July 2015; revision received May 2016; accepted June 2016.

1. Introduction

Peace Arch is the third busiest U.S.-Canada border crossing for passenger automobile traffic, connecting Surrey, British Columbia (Canada) and Blaine, Washington (U.S.). Trucks and commercial vehicles are not allowed to use this crossing, but it still operates 24 hours a day. About 3,500 light-duty vehicles (henceforth "vehicles") pass through it on a slow day, and as many as 4,800 vehicles on a busy day. Waiting times to enter either the U.S. or Canada can reach four hours at certain times of the day. The monetary value of the time spent waiting in queues at U.S.-Canada border crossings is significant. In 2003, the U.S. Department of Transportation reported that the cost of delay while crossing the U.S.-Canada border exceeded $13.2 billion every year (Taylor et al. (2003)). Nguyen and Wigle (2011) pointed out that, while staggering, this figure actually underestimates the cost since it excludes the costs firms incur in maintaining larger inventories as insurance

against late shipments. This poses an important research question: is there an effective workforce policy at security checkpoints that can reduce waiting times without either increasing labor costs or compromising security or customs screening on either side of the border? The goal of this paper is to address this important question by studying traffic flows at Peace Arch: a heavily-used passenger-vehicle crossing. We use publicly available data on northbound traffic from Washington State to British Columbia.

Only recently has the border crossing problem become a topic of study in operations research. Attention has focused on so-called congestion-based workforce policies whereby the number of servers (i.e., the number of active inspection booths in our context) is planned according to the scheduler's expectation of traffic rates, which can be based on historical data or experience. Major border crossing stations often include two stages of inspection. After vehicles complete the first stage, some are randomly selected for a second-stage inspection. Zhang (2009) studied the problem of minimizing delays in the first-stage inspection queue. Zhang et al. (2011) modeled a two-stage security checking system and determined the optimal fraction of passengers selected for second-stage inspection with the objective of balancing security and waiting times. Guo and Zhang (2013) further analyzed the equilibrium state under both no- and partial-information scenarios in a security-checking system equipped with a congestion-based workforce policy. Lin et al. (2014) developed multi-server queueing models to estimate border crossing delays and characterized the transient solutions to the queueing models. They also derived an optimal policy for opening and closing inspection booths over the course of a day, but unlike us they do not consider either workforce scheduling constraints or the potential dependence of service rates on time of day and workload (see below). Since we lack of data on the second-stage inspection, we follow Zhang (2009) in considering only a first-stage queue.

In terms of operations, toll collection on roads is similar in many aspects to border crossings, although delays at border crossings are usually much longer. Most research on toll roads has investigated either tolling for the purpose of pricing congestion or tolling for revenue maximization (see, e.g., Small and Verhoef (2007) and Nagae and Akamatsu (2006)). However, a few studies have examined manpower planning problems at toll collection plazas. Boronico and Siegel (1998) developed a capacity planning analysis for operations at toll collection plazas, and derived the optimal workforce policy. Kim (2009) built a nonlinear integer programming model to study the toll plaza optimization problem, where the cost of waiting times was determined from the steady-state solution of the queueing model. Other studies have discussed the impact of technology on toll plazas. For example, Al-Deek et al. (1997), in a case study, evaluated the benefit of implementing an electronic system (called E-Pass) for toll collection services. Our paper was inspired by

Boronico and Siegel (1998), but our workforce optimization model is significantly more sophisticated because we take into account workforce scheduling constraints and variability in service rate per booth. As demonstrated by our data, both factors are important and statistically significant in the border-crossing setting we study.

Most of the literature mentioned above uses queueing theory to model waiting lines at border crossing stations or toll collection plazas. There is also an extensive literature on how to set up a workforce standard in a heavy-traffic queuing system. Whitt (2007) provides a comprehensive survey of the field. In most of the models in the border-crossing literature, arrival and service rates are time-independent, so that the system can be analyzed at its stationary state. However, in practice, most systems feature significant variability in arrival rates by time of day as is the case at Peace Arch (see Figure 1). Moreover, supply and demand are often not well matched. Figure 1 plots arrival rates (demand) and manpower (supply) over the course of a typical day at Peace Arch. The graph reveals a mismatch between the scheduled workforce (supply) and the arrival rates (demand), which suggests room for improvement. Such a mismatch leads to prolonged waiting times for travelers at peak demand times, and also a waste of staff time when booths are idled. This spurred us to design a new workforce policy that reallocates the number of booths in operation over the course of a day to better accommodate the variable demand while keeping total manpower fixed. We adopt a data-driven approach which forecasts the arrival rate in each hour of a day based on historical demand patterns, and then tries to better match the forecast demand with a feasible workforce schedule.

Our data-driven method also identifies predictors for service variability and uses these predictors in manpower planning. Conventional queueing models often assume that the service rate is either constant or depends only on queue length. However, our data reveals that queue length has little effect on service rate. The strongest predictors of service rate are day of the week, time of day, and the number of active booths. Together, they explain about 45% of the total variance in service rate in our multiple regression model. We incorporate this information into our optimization framework to derive the most efficient and cost-effective work schedule.

The key findings and contributions of this paper are fourfold.

1. We find that the service rate per booth at security checkpoints does not increase with queue length; instead, it is mostly influenced by day of week, time of day and number of active booths. Surprisingly, the service rate decreases as the number of active booths increases. One plausible explanation is that the most able and experienced employees are usually deployed at off-peak hours, whereas to operate all 8 booths less experienced employees need to be

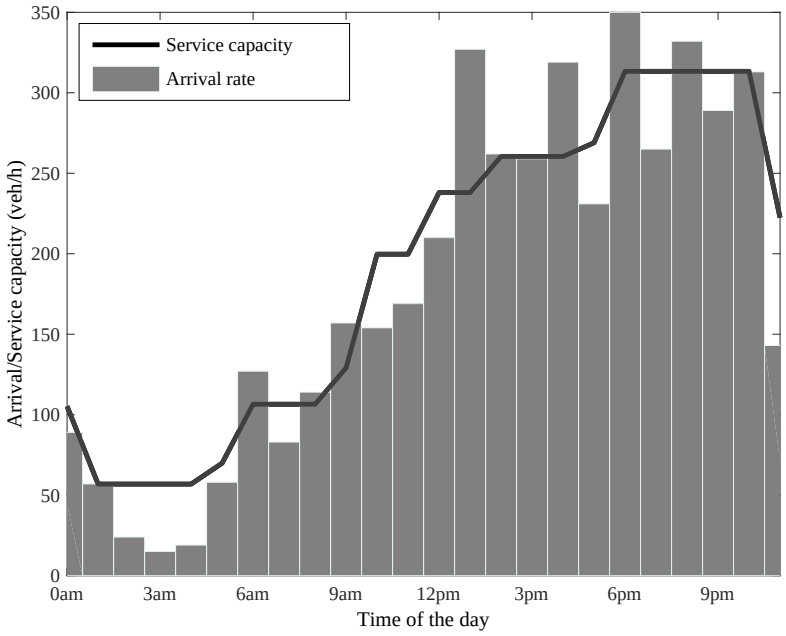


Figure 1 Arrival rate and service capacity in each hour of a day

used. Another explanation is that the waiting area in front of the inspection booths becomes crowded when many booths have been opened, which reduces the speed of the flow.

2. Based on our empirical observations, we propose an integer programming method to obtain an effective workforce policy for each day of the week. The delay function is modeled by a $M_t/G_t/X_t$ queue where X_t is the number of active booths opened in hour t. The key feature of our model is that the service rate (per booth) depends heavily on both t and X_t, and all the parameters in the optimization model are estimated using empirical data. The model can be regarded as an extension of the classical queueing analysis to more realistic situations where service times (per server) are affected by the number of active servers. Thus, albeit motivated by the border-crossing queue, our model and techniques can be widely applied beyond this particular setting.
3. We demonstrate the efficacy of our proposed workforce-policy by benchmarking it against the current policy implemented at Peace Arch. Our simulation test shows that the workforce policy generated using our data-driven approach reduces average waiting time by 17.8% compared to the current workforce policy. This is a substantial improvement given that the two policies have the same total man-hours of labor input.
4. Our results provide insights into the optimal time window to apply congestion pricing. The conventional wisdom in congestion pricing is that tolling is most valuable during peak hours. Nevertheless, in our model, the service capacity can be adjusted according to demand. As a

result, the largest benefits from tolling may not accrue during peak hours, but rather when the servers are most intensively utilized. Because the number of servers is constrained to be an integer value, and the number of staff deployed is constrained by workforce policies, server capacity cannot be matched precisely to demand and peak congestion may occur at off-peak demand times.

The rest of the paper is structured as follows. Section 2 provides a statistical analysis of empirical data collected on service times and arrival rates. Section 3 describes an optimization procedure to devise an effective workforce policy. Section 4 is devoted to numerical simulation of our proposed policy, including a benchmarking of it against the currently implemented policy. Section 5 concludes and identifies some plausible future avenues for research.

2. Empirical Data Analysis

2.1. Data Collection

Government agencies on both sides of the U.S.-Canada border have made serious efforts to quantify border delays and design effective delay-minimizing policies. Transportation authorities currently provide real-time border-crossing delay information to the general public (Cascade 2014), and this provides a fertile environment to investigate related problems with the help of big data. We focus on one of the major U.S.-Canada border-crossing stations, the northbound direction of "Peace-Arch", at which all vehicles except those driven by Nexus-pass holders entering Canada from the U.S. are inspected. The raw data available on Cascade (2014) includes the *arrival rate*, *departure rate*, *number of active booths*, and *vehicles in queue*. *Arrival rate* counts the number of vehicles, in five-minute segments, passing through the first detector to join the inspection queue. Typical arrival rates are 20-25 vehicles per five minutes during the day, and less than five vehicles per five minutes during the night. *Departure rate* counts the number of vehicles, in five-minute segments, passing through the first detector that all vehicles have to traverse when leaving the station. Typical service rates per booth are five vehicles per five minutes. *Number of active booths* refers to the number of booths in operation. Since each active booth has to be staffed by a security employee, the number of active booths determines the staffing level required at a given time. *Vehicles in queues* counts the total number of vehicles currently waiting in the first-stage inspection queue. We use the border-crossing data from April 1, 2014 – June 30, 2014 as the training data to develop a regression model for the service and arrival rates. We use the data from September 1, 2014 – October 30, 2014 as test data to evaluate the workforce policy we propose. We intentionally leave a gap between the training and test data to mimic the realistic scenario when the training data may not be updated immediately due to delays in collecting and pre-processing the data.

2.2. Empirical Distribution of Service Times

To develop a regression model for the service rate, we only use observations with queue length greater than 20. The reason is that queues longer than 20 vehicles are unlikely to be completely served in a five-minute window, and hence the departure rate can be used as a proxy for the service rate. We then search for factors that may contribute to variation of the service rate, including day of week, time of day, number of active booths, and queue length. A multiple linear regression model is proposed as follows:

$$\text{service-rate} = \beta_0 + \beta_1 \cdot \text{day} + \beta_2 \cdot \text{time} + \beta_3 \cdot \text{no.-of-active-booths} + \beta_4 \cdot \text{queue-length} + \epsilon,$$

where *service rate* is our response variable, $\beta_0, \ldots, \beta_4$ are regression coefficients and ϵ is zero-mean random noise. The regression model has $R^2 = 45.1\%$ (see Figure 8 in the appendix) and all explanatory variables are statistically significant at the 0.001 level. Although respectable, the goodness-of-fit might be improved by including individual traveler characteristics, such as previous entry record, nationality, and other personal information. However, this sensitive information is not available on the website for the general public, and thus unavailable to us.

The regression analysis also reveals how each explanatory variable influences the service rates (see Table 1). In particular,

1. Service rates reach a peak in the morning, decrease to a minimum around noon, and recover in the evening.
2. Service rates are, in general, higher on Monday–Thursday than on other days.
3. Service rates decrease with the number of active booths (or equivalently, the staffing level).
4. There is no strong evidence that service rates increase when the queue is long, which suggests that the border guards are doing their jobs properly and are not rushed when queues are long.

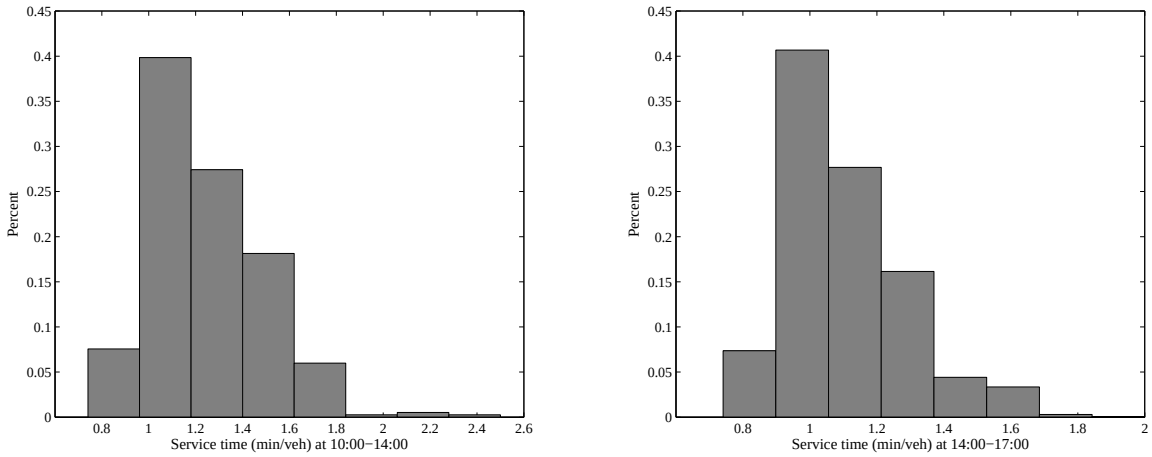
In order to explain the above observations, we carried out a literature review and also interviewed staff members at the security check station. One possible reason for the slow service rate around noon is that this is a peak time for Canadians shopping in the U.S. to return to Canada, and it takes staff additional time to inspect their purchases. Travelers in the morning or evening, and on Monday–Thursday, are more likely to cross the border for work/business purposes, and inspection is usually shorter for them. When there are more active booths, the service rate per booth decreases. One possible reason is that the waiting area in front of the booths becomes very crowded which can impede motorists from driving through the booth. Another possibility is that the most able and experienced employees are usually deployed at off-peak hours. At peak times, less experienced employees need to be used.

Table 1 How different factors influence the service rate

		Day	**Time**
		Friday	12:00–13:00
		Sunday	11:00–12:00
		Saturday	10:00–11:00
		Thursday	13:00–14:00
		Monday	22:00–23:00
	Slowest	Wednesday	19:00–20:00
		Tuesday	18:00–19:00
		Number of Active Booths	21:00–22:00
		8	23:00–0:00 (+1 day)
		6	17:00–18:00
		7	20:00–21:00
Service Rate per Booth		3	14:00–15:00
		5	16:00–17:00
		4	15:00–16:00
	↓	2	9:00–10:00
		1	0:00–1:00
		Vehicles in Queue	2:00–3:00
		(150,200]	8:00–9:00
		(200,250]	1:00–2:00
		(250,300]	4:00–5:00
		(300,600]	3:00–4:00
	Fastest	(100,150]	7:00–8:00
		(50,100]	6:00–7:00
		(20,50]	5:00–6:00

To assess the independent contributions of the explanatory variables, we ran a simple linear regression for service rate with respect to each of the four explanatory variables separately. The results (see Figure 9 in the appendix) indicate that 40% and 15% of the total variance of service rate can be explained by time of day and number of active booths, respectively. The predictive powers of day of week and vehicles in queue are less significant (with $R^2 \leq 10\%$). For this reason, in our data-driven method we only model the effects of time of day and number of active booths.

In order to formulate the staffing problem using mathematical programming, we need to approximate the service time distribution using a certain parametric form. After extensive testing of various parametric forms using maximum likelihood estimation, we found that a Gamma distribution gave the best fit. For each hour i and for each number of active booths x_i, we run a goodness-of-fit test for the Gamma distribution and estimate its mean mean $m_i(x_i)$ and variance $\sigma_i^2(x_i)$. The p-values for the goodness-of-fit test range between 0.4 and 0.9, which supports the null hypothesis that the fit is accurate. The estimated parameters $m_i(x_i)$ and $\sigma_i^2(x_i)$ are summarized in Table 5 in the appendix. Figure 2 presents histograms for service-time distributions in two time periods (10:00–14:00 and 14:00–17:00, each with four active booths).



(a) Histogram of service time during 10:00–14:00 (b) Histogram of service time during 14:00–17:00

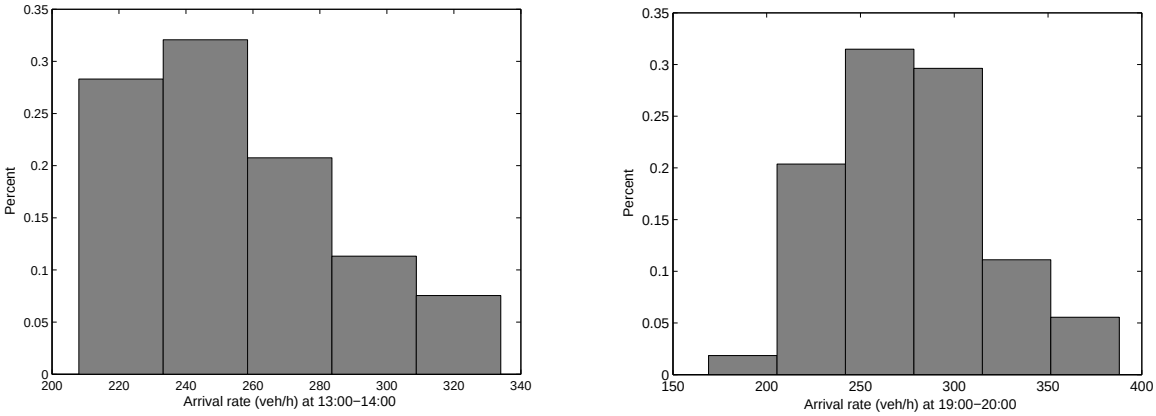
Figure 2 Distributions of service times with four active booths

2.3. Arrival Process

We assume that the customer arrival process can be modeled as a non-homogeneous Poisson process that depends on the day of week. The arrival patterns on Tuesday, Wednesday, and Thursday are actually very similar, so we treat Tuesday/Wednesday/Thursday (T/W/T) as the same day but fit the arrival rates in different days with different distributions. Without loss of generality, the subsequent discussion on fitting the arrival process assumes that the day is T/W/T.

Let $N(t)$ be the Poisson counting function for $t = 0, 1, \ldots 24$; i.e., $N(t)$ denotes how many customers arrive between hour 0 and hour t. We test the hypothesis "customer arrivals within one hour follow a Poisson distribution with the parameter from maximum likelihood estimation" using data on intra-day arrival rates by testing the following two properties: (a) $N(t+1) - N(t)$ follows a Poisson distribution with some rate λ_t; (b) independent increments.

In order to test property (a), we fit the arrival-rate in the training data using a non-homogeneous Poisson process, whose intensity function $\lambda(t)$ is assumed to be constant during each one-hour window, e.g., 7:00am–7:59am, 8:00am–8:59am, but can change on the hour. The p-value of the goodness-of-fit test is $> 5\%$, which supports property (a). We test property (b) by examining the empirical joint distribution of arrival rates across different hours, and in call cases the results support property (b). Having tested properties (a) and (b), we can use the non-homogeneous Poisson process to approximate the intra-day arrival process.



(a) Arrival rate at 13:00–14:00 on T/W/T (b) Arrival rate at 19:00–20:00 on T/W/T

Figure 3 Two examples of hourly arrival rates

Two examples of hourly arrival rates, 13:00–14:00 and 19:00–20:00, are depicted in Figure 3. As reported in Table 2, goodness-of-fit tests are satisfied for both distributions. Table 2 summarizes the fitted values of arrival rates.

Table 2 Goodness-of-fit tests for the arrival processes on a T/W/T

Goodness of fit at 13:00–14:00 (lambda=247)						
Arrival rate	(210, 230)	(230, 250)	(250, 270)	(270, 290)	(290, 330)	
obsCounts	8	13	12	10	10	
expCounts	5.9394	12.9995	14.2259	10.3786	9.0715	
chi2stat	1.172		**p-value**	0.7597		
Goodness of fit at 19:00–20:00 (lambda=278)						
Arrival rate	< 245	(245, 267)	(267, 289)	(289,311)	(311,333)	>333
obsCounts	7	9	13	9	9	7
expCounts	9.2789	9.7159	10.3099	9.1644	7.8871	6.7884
chi2stat	2.2927		**p-value**	0.6821		

Table 3 Summary of hourly arrival rates

Hours	1	2	3	4	5	6
Arrival Rates	110.65	51.25	27.10	19.90	23.25	70.00
Hours	7	8	9	10	11	12
Arrival Rates	118.58	84.83	116.68	156.48	170.00	196.80
Hours	13	14	15	16	17	18
Arrival Rates	212.85	247.28	274.13	287.73	292.15	276.78
Hours	19	20	21	22	23	24
Arrival Rates	260.15	277.79	282.95	278.63	239.70	171.20

2.4. Active Booths

We analyzed the *active booths* column of the training data. On most days, the staffing level followed a sinusoidal-like curve, reaching its minimum at about 3 a.m., increasing until late afternoon, and then decreasing until reaching a minimum at about 3 a.m. of the next day. Based on our conversations with staff members at the security check station, the underlying reason for the sinusoidal-like staffing curve is that most staff members prefer to work for a single continuous period of time at the inspection booths without any major interruption (a short break is possible), so they can work on some other desktop duties inside the station for the remainder of their work hours. As a result, the staffing level curve cannot have multiple oscillations (see Figure 1). The data also indicate that no more than eight booths are ever active even though the station has ten booths. The total number of man hours during a day is also relatively fixed (between 80 and 95).

3. Optimizing the Staffing Level

Since the arrival rate depends on the day of the week, the optimal staffing schedule generally does too, and so do the staffing schedules in our data. However, since the distributions of service and arrival rates were similar for Tuesday, Wednesday, and Thursday, we will use the same schedules for these days. In this section, we focus on the T/W/T schedule of every week. The method used can also be applied to determining staffing schedules for other days of the week.

3.1. Notation and assumptions

The notation and assumptions used in the balance of the paper are listed below.

Decision variables

x_i	Number of servers during the hour period i.

Parameters

$Q_i(x_i)$	Mean queue length with x_i servers (booths) during hour period i;
λ_i	Arrival rate during hour period i;
$m_i(x_i)$	Mean service time per booth with x_i servers (booths) during hour period i;
$\rho_i(x_i)$	Utilization factor with x_i servers (booths) during hour period i;
C_A	Coefficient of variation of inter-arrival times;
$C_S(x_i)$	Coefficient of variation of service times with x_i servers;
$\sigma_i(x_i)$	Standard deviation of service times during hour period i with x_i servers;
K	Upper limit of total man hours on that day

Assumptions

(i) Any change in staff occurs at the beginning of each hour.
(ii) Each staff person works for at least one full hour.
(iii) Both service rates and arrival rates change only at the beginning of each hour.

3.2. Optimization Model

The staffing problem is to determine the optimal number of security check personnel as a function of time. The goal is to minimize travelers' expected waiting times given a fixed total number of man hours. In the following, we formulate an optimization model based on $M/G/X$ queues to obtain an effective schedule for each day.

By the Pollaczek-Khinchine approximation formula for the $M/G/x$ queue (see Pollaczek (1930) and Khinchin (1967)), the mean queue length at time i is

$$Q_i(x_i) = \frac{\rho_i(x_i)\sqrt{2(x_i+1)}}{1-\rho_i(x_i)} \frac{C_A^2 + C_S^2(x_i)}{2}, \tag{1}$$

where $\rho_i(x_i)$ is the system utilization factor given by

$$\rho_i(x_i) = \frac{\lambda_i m_i(x_i)}{x_i}.$$

Note that the original Pollaczek-Khinchine formula was developed for $M/G/1$ queues, but later extended to $M/G/x$ queues and used extensively in practice (see, e.g., Cachon and Terwiesch (2009)). Whitt (1993) described this approximation as "usually an excellent approximation, even given extra information about the service-time distribution." This result is also known as Kingman's law of congestion (see Gans et al. (2003)).

Since the inter-arrival times are modeled as exponential random variables, it is clear that their coefficient of variation is $C_A = 1$. The coefficient of variation of general service times with x_i servers is given by

$$C_S(x_i) = \frac{\sigma_i(x_i)}{m_i(x_i)}.$$

Then (1) reduces to

$$Q_i(x_i) = \frac{\rho_i(x_i)\sqrt{2(x_i+1)}}{1-\rho_i(x_i)} \frac{1}{2}\left(1 + \frac{\sigma_i^2(x_i)}{m_i^2(x_i)}\right). \tag{2}$$

Our objective is to minimize the expected waiting times (since the service time is much smaller than the waiting time, we will use waiting time instead of total sojourn time as the objective) of all passengers who cross the border on that day. By Little's Law, this can be expressed as

$$\frac{\sum_{i=1}^{24} Q_i(x_i)}{\sum_i \lambda_i}. \tag{3}$$

The constraints for this optimization model include the following: (a) The maximum number of active booths is eight; (b) the staffing level curve must monotonically increase from time x_{min} to time x_{max}, and monotonically decrease from time x_{max} to x_{min} (of the next day), where x_{min} and x_{max} denote the times when the staffing level reaches the minimum and maximum value in a day, respectively; (c) total man hours cannot exceed K, where K ranges from 80 to 95. All of these constraints are explained in Section 2.4. These are the main practical constraints when it comes to workforce scheduling at the booths. In addition, the schedule for each employee is preferred to be stable from week to week; each employee needs to take at least two days off in a week, at least one of which has to be on the weekend. However, since we study the scheduling problem at the hourly level in T/W/T, those constraints are not relevant to our model and hence will not be incorporated in our analysis.

$$\min_{x_i} \quad \frac{\sum_{i=1}^{24} Q_i(x_i)}{\sum_{i=1}^{24} \lambda_i} \tag{4}$$

$$\text{s.t.} \quad 1 \le x_i \le 8, \qquad \forall i = 1, \ldots, 24; \tag{5}$$

$$x_{\text{mod}(i-1+x_{min},24)} \le x_{\text{mod}(i+x_{min},24)} \qquad \forall i = 1, 2, \ldots, \text{mod}(x_{max} - x_{min}, 24); \tag{6}$$

$$x_{\text{mod}(i-1+x_{max},24)} \le x_{\text{mod}(i+x_{max},24)} \qquad \forall i = 1, 2, \ldots, \text{mod}(x_{min} - x_{max}, 24); \tag{7}$$

$$\sum_{i=1}^{24} x_i \le K, \qquad \forall i = 1, \ldots, 24; \tag{8}$$

$$x_i \in \mathbb{Z}, \qquad \forall i = 1, \ldots, 24. \tag{9}$$

Constraints (5), (6), and (8) correspond to the aforementioned constraints (a), (b), and (c), respectively. Constraint (9) forces x_i, the staffing level in each hour, to be integer-valued. We have to search for the pair (x_{min}, x_{max}) which yields the minimum objective value. Since there are only 24^2 possible combinations, an exhaustive search for (x_{min}, x_{max}) is tolerable.

3.3. Implication for Congestion Pricing

So far we have focused on how to reduce congestion at the border crossing by a more efficient allocation of service capacity. An alternative way to reduce congestion is to smooth out demand by imposing tolls at certain times of a day. Most of the literature on congestion pricing has assumed that tolls are levied during peak hours (e.g. Small and Verhoef (2007)). This is reasonable for roads, airports, and other facilities with a capacity that is constant throughout the demand cycle. However, at border crossings, service capacity can be adjusted hourly, and it is not obvious whether congestion is worse when demand is high and many booths are open, or when demand is low and

only a few booths are active. The congestion externality is highest when a marginal increase in the arrival rate causes the largest increment in total waiting time. The incremental time can be solved by computing the gradient of the objective value of the integer programming problem (4)-(9) with respect to the arrival-rate vector $(\lambda_i)_{i=1,\ldots,24}$. We plot the gradient in Figure 4 and compare the curve with the arrival-rate vector on a typical day. According to the plot, the increment in average waiting time has two peaks: one at 9am and the other from 8-10pm. By contrast, the arrival rate peaks around 3pm. The discrepancy between the two curves can be explained by the difference in service rates and number of active booths across different hours of a day.

Additional information would be required to derive an optimal toll schedule for the border crossing at Peace Arch. One is the value of travelers' time. A second is the demand curve for crossings by time of day which depends on the scope for travelers to reschedule trips, use alternative crossings, and/or avoid taking trips altogether. Third, it would be necessary to know how diversion of trips to other crossings would affect queuing delays there. Tolling the Peace Arch will exacerbate congestion delays at other crossings if they remain untolled. Indeed, depending on the level of the tolls and the degree of substitutability between crossings, the benefits could be negative (Small and Verhoef (2007), Section 4.2.1). Since none of this additional information is available, we do not attempt to derive optimal tolls in this study.

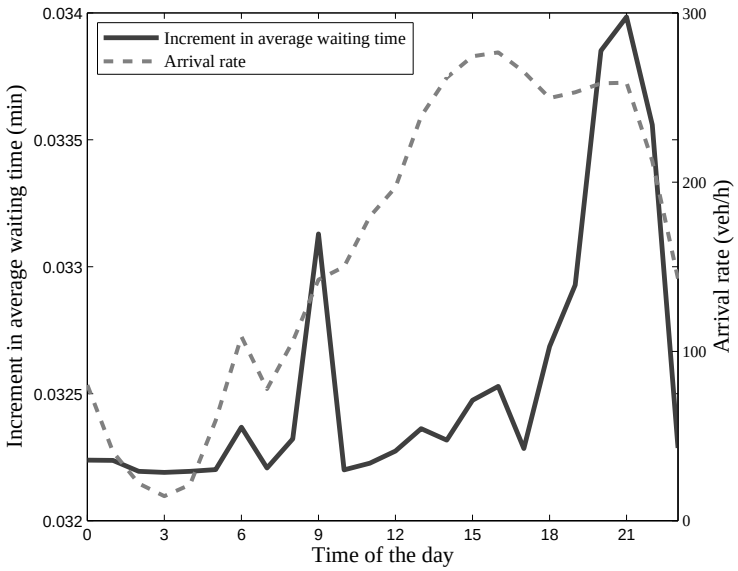


Figure 4 Incremental waiting times

These challenges notwithstanding, tolling Peace Arch and/or other U.S.-Canada border crossings is a distinct possibility. Twelve of the bridges or tunnels linking Ontario and the U.S. are tolled.

Within Metro Vancouver, the Port Mann Bridge and Golden Ears Bridge are tolled, and planned replacements for the Pattullo Bridge and the George Massey Tunnel are to be tolled as well. Congestion pricing of roads in Canada received support in a recent study by a group of eminent Canadian economists (see Ecofiscal-Commission (2015)). Tolling border crossings would not only help to reduce queueing delays, but also reduce air pollution and greenhouse gas emissions. It would generate revenues for cash-strapped governments in British Columbia and other provinces. Tolling is consistent with the user-pay principle that users of a service (i.e., border crossings) should help to cover its costs. Tolls would also discourage cross-border shopping that often emerges as a public issue in Canada when the U.S.-Canada exchange rate is favorable to Canadian shoppers. Indeed, in June 2014, a professor at the University of British Columbia suggested tolling border crossings at a time when the Canadian dollar was worth about USD $0.90.

3.4. Other Types of Passengers

Other types of vehicles and users that regularly commute across the U.S.-Canada border include buses, trucks, FAST card holders, and NEXUS pass holders. Only a few major ports can accommodate passengers of those types. Instead of going through the booths, buses have to park and disembark all passengers, who have to go through the security check inside office buildings. Trucks have to go through security inspection at different booths than the cars. Also, the procedure and skills for inspecting trucks are different from those for cars. Thus, trucks and vehicles are usually not served by the same pool of personnel, and we could consider border crossings of vehicles and trucks as two independent systems. However, since the traffic volume of trucks is much smaller than vehicles, most border crossings open at most two active booths dedicated for trucks. Thus, the manpower planning problem for trucks is less complicated, and is not considered here. Commercial drivers who are regarded as lower risk can apply for a FAST card, which allows them to go through a dedicated booth with expedited inspection service. Similar to trucks, it suffices to serve all the FAST card holders with a single active booth, so the manpower planning problem is also straightforward for this category of travellers.

The last category, Nexus pass holders, are pre-approved low-risk travelers, who can pass the entry point using automated self-serve kiosks without inspection. Currently, about $30\% - 40\%$ of the daily traffic crossing the Peace Arch or Pacific (excluding trucks and buses) consists of Nexus pass holders. With this proportion increasing in the future, the prolonged waiting at border crossings could be alleviated because the Nexus pass holders requite little service capacity. We briefly explore such a future here by assuming that a proportion α of the vehicles hold Nexus passes and are diverted to the Nexus lanes. We calculate the corresponding average waiting times under

staffing schedules generated using our data-driven method, and plot the curve in Figure 5. We find that issuing only 3% of the current vehicle passengers with Nexus passes reduces waiting time by about 50%. This indicates that a modest investment in expanding Nexus penetration could lead to a significant reduction in the waiting time provided the reduced waiting time does not induce significant latent demand from either additional travel or retiming of trips.

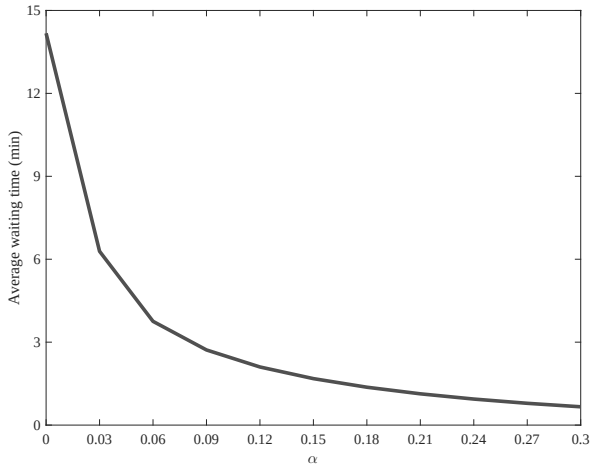


Figure 5 **Average waiting time corresponding to an $\alpha \times 100\%$ reduction in (non-Nexus) vehicle arrival rate**

4. Simulation

In this section, we apply the workforce policy derived in Section 3 to the test dataset which covers a two-month period from September 1, 2014 to October 31, 2014. We first calculate the average waiting-time in the current practice using Little' law (3), in which the real-time queue-length information can be inferred from the online records. We then solve the optimal workforce policy using the integer program problem (4)-(9), in which $m_i(x_i)$ and $\sigma_i(x_i)$ $(i = 1, 2, \ldots, 24)$ are predicted using the regression model developed in Section 2.2, and λ_i $(i = 1, 2, \ldots, 24)$ are estimated based on the online records. We compute average waiting-time as the objective of the integer programming problem, and compare it to that in the current practice on T/W/T in each week during the test period. The results are listed in Table 4. The last column, reduction ratio, is computed using the formula

$$\text{reduction ratio} = \frac{\text{avg. waiting time under current policy} - \text{avg. waiting time under our policy}}{\text{avg. waiting time under current policy}} \quad (10)$$

Table 4 shows that on most days the reduction ratio is positive, which means that our policy reduces the average waiting time on most days. Over a full day, on average (not weighted by traffic

volume in each day) waiting time is reduced by 17.8%. This is a significant improvement considering that our policy only reallocates the workforce without adding any extra manpower. Because the proposed workforce policy is derived using predicted arrival rates, it may not be the optimal policy. Therefore, sometimes it can perform worse than the current policy as Table 4 shows.

Date	Current	Our Policy	Reduction Ratio (%)
9.2	14.9	14.1	5.2
9.3	19.1	14.2	25.6
9.4	14.0	13.4	3.8
9.9	24.0	12.4	48.3
9.10	12.4	14.1	-13.5
9.14	17.0	13.7	19.5
9.16	14.8	13.3	10.1
9.17	18.8	16.5	12.3
9.18	21.3	16.6	22.4
9.23	51.6	16.2	68.6
9.24	18.3	17.1	6.6
9.25	17.0	17.5	-2.7
9.30	20.7	18.9	8.5
10.1	17.7	14.9	16.2
10.2	13.8	15.1	-9.5
10.7	15.3	14.9	2.2
10.8	30.7	17.4	43.2
10.9	15.5	15.9	-2.9
10.14	16.1	16.8	-4.6
10.15	15.3	16.3	-6.6
10.16	16.9	16.9	0.1
10.21	15.5	16.1	-3.8
10.22	19.9	15.3	23.2
10.28	13.2	11.8	10.1
10.29	19.9	19.9	-0.3
10.30	16.3	13.3	18.4
average	18.8	15.5	17.8

Table 4 Comparison of average waiting times between two staffing policies (mins)

Figure 6 plots a staffing level curve derived using our data-driven method, and a typical staffing level curve based on the current schedule. The figure reveals that using the data-driven method allows more staff members to be scheduled in the morning, and fewer in the evening, thus supporting a steadier number of active booths over the course of the day. One reason for operating more booths in the morning is that queues that develop early in the day could persist until late in the evening – thereby resulting in large cumulative delays for travelers and high costs in wasted time and inconvenience. Hall (1991) p.220 explains why scheduling additional capacity early enough to prevent queueing is a good strategy. Specifically, he argues that "waiting until a queue becomes a significant problem before adding capacity is a very risky policy because it may be impossible to catch up with demand". While his demonstration applies to situations in which demand varies predictably, the logic still applies when demand is random as long as the random variations do not dominate the predictable variations.

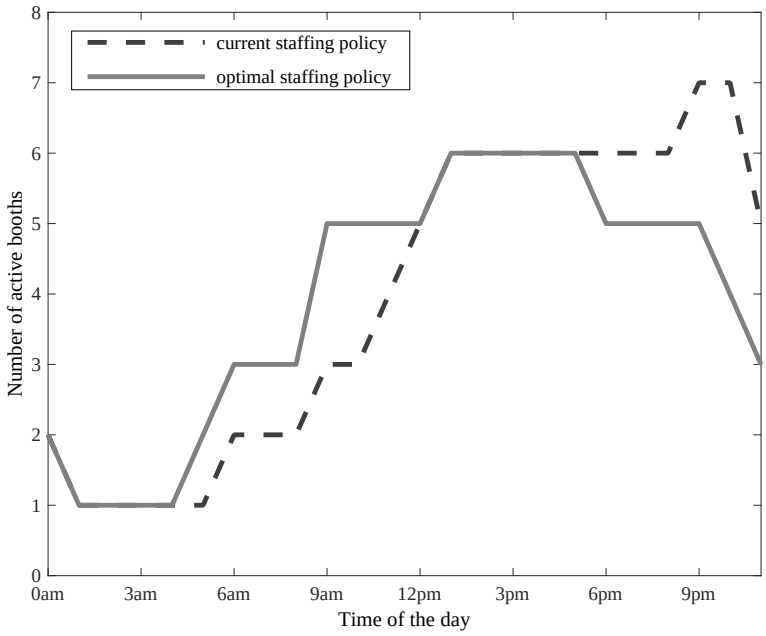


Figure 6 **Comparison of staffing levels between two staffing policies**

So far we have focused on minimizing the expected number of vehicles in the queue, or equivalently, the average waiting time of a vehicle. Another important performance metric often used in labor staffing and scheduling is the service level (see, e.g., Thompson (1997)), which is measured as the percentage of customers getting served within a specified time target. Because the service level in an M/G/m queue does not admit a closed-form expression, we cannot directly incorporate the service level constraints into our integer programming. However, we can compute the service level in each hour under a given staffing schedule. Figure 7 compares service levels (with respect to a 15-minute target) in each hour using the schedule solved from our data-driven method and the schedule plotted in Figure 6, where the reduction ratio follows the same definition as in (10). Figure 7 shows that our staffing policy maintains 70% service levels for most hours while the current staffing policy suffers from extremely low service levels during the 4am-10am time window. Our service levels are slightly worse during the 5pm-10pm time window, but this is expected since we are not adding more manpower. Overall, our staffing policy gives rise to much better and more stable service levels compared to the current staffing policy.

5. Conclusion and Future Directions

In this paper, we propose a data-driven approach to optimizing the work-schedule of security check staff at border crossing stations. We identify several variables that help to predict traffic rates and

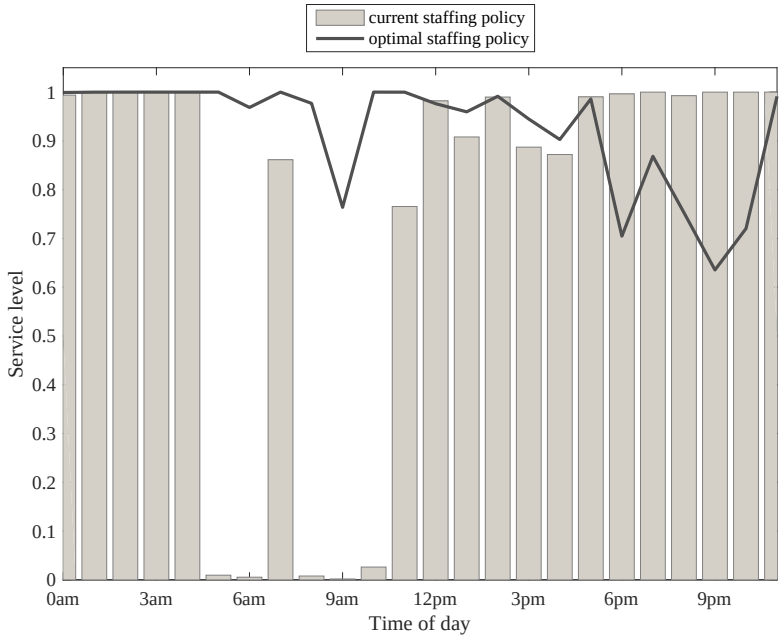


Figure 7 Comparison of service levels between the two staffing policies

service rates, including day of week, hour of day, and number of active booths. This finding enables us to design the optimal work schedule to minimize average vehicle waiting times. The formulation of the optimization problem considers practical constraints, such as not disrupting a staff member's working hours. Our simulation using real border crossing data shows that our manpower policy significantly improves on the current schedule with respect to vehicle waiting times.

This work offers insights into general service operations. When demand is time-varying, it is important to forecast demand and service rates so that capacity can be matched commensurately with demand. We proposed a data-driven method to forecast the demand and the capacity at different times and under different conditions. Since the adjustment is usually subject to certain constraints, a natural approach is to formulate the problem as a mathematical program in which the constraints can be expressed either explicitly or approximately.

The method can be adapted to many other capacity allocation problems in transportation and other economic sectors. Nexus lanes at border crossings are one obvious instance. However, most border crossings have only one lane dedicated to Nexus so that the choice of capacity effectively reduces to a yes-or-no decision. Moreover, although the Nexus program has existed for over a decade it has still had only a limited impact on total border crossing delays. According to information reported in Seghetti (2014), less than 2.5 percent of land entries to the U.S. in Fiscal Year 2013 were admitted under Nexus. For Canadians, the application process for a Nexus card takes about

a month, and applicants must arrange interviews with officers of the Canada Border Services Agency. Various restrictions on usage of Nexus lanes apply, and at six of the 20 land border crossings between Canada and the U.S. the Nexus lanes are open three hours per day or less (See http://www.cbsa-asfc.gc.ca/prog/nexus/land-terre-eng.html).

Airport security is another potential application that has been studied by Aksin et al. (2007) and Casado et al. (2006). Further applications are allocating staff to motor vehicle licensing or passport renewal stations, clerks to supermarket counters, waiters to restaurant tables and so on. The time pattern of demand can vary substantially across these and other settings. As shown in Figure 1, the demand profile at Peace Arch is unimodal with a peak in the late afternoon or early evening. At other border crossings, the profile may be bimodal due to morning and afternoon commuting peaks. At facilities such as restaurants, it could be trimodal with peaks associated with mealtimes. The general methodology described in the paper can readily be adapted to accommodate these variations. Another consideration is how employee productivity is affected by workload. The data we use indicate that employee service rates at Peace Arch are not affected by queue length. However, employee output can be affected by workload in other settings. Some studies have found that high workloads induce employees to speed up, while other studies found that employees slow down (see Tan and Netessine (2014)).

A number of topics remain for future research. For example, we do not consider the fact that some vehicles are selected to pass a second-stage inspection and that the two stages actually share the same service resources (i.e., security check staff members). We do not incorporate this feature into our model because we lack of data on the second-stage inspection. Another limitation is that our model predicts service and arrival rates on a regular weekday or weekend. It may not be applicable on days with special events such as a large sports or entertainment event in a city near the border. To investigate the workforce policy on these days, alternative data and forecasting method are required. Third, our method focuses on allocating service resources at different times at a single border crossing station. Another interesting problem is how to allocate staff members across different border crossing stations. This is practical for border crossing stations that are very close to each other such as Peace Arch and Pacific in British Columbia, and the bridges crossing the Niagara River between New York State and Ontario.

Acknowledgments

The authors are grateful to Professor Yinyu Ye for his valuable suggestions on formulating the staff level optimization problem. The research was partially conducted when Mengqiao Yu participated in the MITAC Globalink exchange program at Sauder School of Business, University of British Columbia, and Mengqiao

Yu was fully supported by the Mitacs-Globalink program during that time. The research of Yichuan Ding is partially supported by NSERC PGPIN 436156-13. The research of Robin Lindsey is partially supported by SSHRC grant 435-2014-2050. The research of Cong Shi is partially supported by NSF grants CMMI-1362619 and CMMI-1451078.

References

Aksin, Z., M. Armony, V. Mehrotra. 2007. The modern call center: A multi-disciplinary perspective on operations management research. *Production and Operations Management* **16**(6) 665–688.

Al-Deek, H. M., A. A. Mohamed, A. E. Radwan. 1997. Operational benefits of electronic toll collection: case study. *Journal of Transportation Engineering* **123**(6) 467–477.

Boronico, J. S., P. H. Siegel. 1998. Capacity planning for toll roadways incorporating consumer wait time costs. *Transportation Research Part A: Policy and Practice* **32**(4) 297–310.

Cachon, G., C. Terwiesch. 2009. *Matching supply with demand*, vol. 2. McGraw-Hill Singapore.

Casado, S., J. A. P. Bonrostro, M. Laguna. 2006. Diseño de un sistema para la resolución del problema de programación de turnos en un aeropuerto. *Estudios de economía aplicada* **24**(2) 641–642.

Cascade. 2014. Cascade gateway border data warehouse. http://www.cascadegatewaydata.com/.

Ecofiscal-Commission. 2015. We cant get there from here: Why pricing traffic congestion is critical to beating it. http://ecofiscal.ca/reports/traffic/.

Gans, N., G. Koole, A. Mandelbaum. 2003. Commissioned paper: Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management* **5**(2) 79–141.

Guo, P., Z. G. Zhang. 2013. Strategic queueing behavior and its impact on system performance in service systems with the congestion-based staffing policy. *Manufacturing & Service Operations Management* **15**(1) 118–131.

Hall, R. W. 1991. *Queueing methods: for services and manufacturing*. Prentice-Hall international series in industrial and systems engineering, Prentice Hall, Englewood, Cliffs, NJ.

Khinchin, A. Y. 1967. The mathematical theory of a stationary queue. Tech. rep., DTIC Document.

Kim, S. 2009. The toll plaza optimization problem: Design, operations, and strategies. *Transportation Research Part E: Logistics and Transportation Review* **45**(1) 125 – 137.

Lin, L., Q. Wang, A. W. Sadek. 2014. Border crossing delay prediction using transient multi-server queueing models. *Transportation Research Part A: Policy and Practice* **64**(0) 65 – 91.

Nagae, T., T. Akamatsu. 2006. Dynamic revenue management of a toll road project under transportation demand uncertainty. *Networks and Spatial Economics* **6**(3-4) 345–357.

Nguyen, T. T., R. M. Wigle. 2011. Border delays re-emerging priority: Within-country dimensions for canada. *Canadian Public Policy* **37**(1) 49 – 59.

Pollaczek, F. 1930. Über eine aufgabe der wahrscheinlichkeitstheorie. i. *Mathematische Zeitschrift* **32**(1) 64–100.

Seghetti, L. 2014. Border security: Immigration inspections at port of entry. *Current Politics and Economics of the United States, Canada and Mexico* **16**(2) 157–204.

Small, K. A., E. T. Verhoef. 2007. *The economics of urban transportation*. Routledge.

Tan, T. F., S. Netessine. 2014. When does the devil make work? an empirical study of the impact of workload on worker productivity. *Management Science* **60**(6) 1574–1593.

Taylor, J. C., D. Robideaux, G. C. Jackson. 2003. The u.s.-canada border: Cost impacts, causes, and short to long term management options. Technical Report, Grand Valley State University, Allendale, Michigan.

Thompson, G. M. 1997. Labor staffing and scheduling models for controlling service levels. *Naval Research Logistics* **44**(8) 719–740.

Whitt, W. 1993. Approximations for the GI/G/m queue. *Production and Operations Management* **2**(2) 114–161.

Whitt, W. 2007. What you should know about queueing models to set staffing requirements in service systems. *Naval Research Logistics* **54**(5) 476 – 484.

Zhang, Z. G. 2009. Performance analysis of a queue with congestion-based staffing policy. *Management Science* **55**(2) 240–251.

Zhang, Z. G., H. P. Luh, C.-H. Wang. 2011. Modeling security-check queues. *Management Science* **57**(11) 1979–1995.

Appendix. Additional Tables and Figures.

SAS Procedure

Dependent Variable: service_rate service rate

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	42	51.5993255	1.2285554	100.96	<.0001
Error	5171	62.9262891	0.0121691		
Corrected Total	5213	114.5256146			

R-Square	Coeff Var	Root MSE	service_rate Mean
0.450548	12.64687	0.110314	0.872259

Source	DF	Type I SS	Mean Square	F Value	Pr > F
day	6	4.84126989	0.80687832	66.31	<.0001
time	23	45.38059661	1.97306942	162.14	<.0001
booth_number	7	0.69496543	0.09928078	8.16	<.0001
queue_level	6	0.68249356	0.11374893	9.35	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
day	6	3.32255301	0.55375883	45.51	<.0001
time	23	26.47068792	1.15089947	94.58	<.0001
booth_number	7	0.69177158	0.09882451	8.12	<.0001
queue_level	6	0.68249356	0.11374893	9.35	<.0001

Figure 8 **R-square for a multiple regression with $Y =$ service rates, and $X_1 =$ day, $X_2 =$ time, $X_3 =$ active booth number, $X_4 =$ vehicles in queue.**

Table 5 Parameters for service time distributions (with the number of active booths x_i)

i=1,2,3,4,5			i=6,7,8,9,10		
x_i	$m_i(x_i)$	$\sigma_i^2(x_i)$	x_i	$m_i(x_i)$	$\sigma_i^2(x_i)$
1	1.055	0.614	1	0.859	0.387
2	1.14	0.429	2	1.127	0.3
3	1.048	0.321	3	1.395	0.249
4	1.142	0.245	4	1.663	0.212
5	1.158	0.185	5	1.931	0.184
6	1.105	0.137	6	2.199	0.161
7	1.097	0.096	7	2.467	0.142
8	1.126	0.06	8	2.735	0.125
i=11,12,13,14			i=15,16,17		
x_i	$m_i(x_i)$	$\sigma_i^2(x_i)$	x_i	$m_i(x_i)$	$\sigma_i^2(x_i)$
1	1.025	0.364	1	0.994	0.332
2	1.084	0.291	2	1.034	0.267
3	1.143	0.248	3	1.073	0.229
4	1.202	0.217	4	1.113	0.202
5	1.261	0.193	5	1.152	0.181
6	1.32	0.174	6	1.191	0.163
7	1.378	0.158	7	1.231	0.149
8	1.437	0.144	8	1.27	0.136
i=18,19,20,21,22,23,24					
x_i	$m_i(x_i)$	$\sigma_i^2(x_i)$			
1	0.765	0.343			
2	0.936	0.268			
3	1.046	0.224			
4	1.079	0.193			
5	1.115	0.169			
6	1.149	0.149			
7	1.184	0.133			
8	1.203	0.118			

SAS procedure-Day

Dependent Variable: service_rate service rate

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	4.8412699	0.8068783	38.30	<.0001
Error	5207	109.6843447	0.0210648		
Corrected Total	5213	114.5256146			

R-Square	Coeff Var	Root MSE	service_rate Mean
0.042272	16.63922	0.145137	0.872259

SAS procedure-Time

Dependent Variable: service_rate service rate

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	23	46.3093808	2.0134513	153.19	<.0001
Error	5190	68.2162339	0.0131438		
Corrected Total	5213	114.5256146			

R-Square	Coeff Var	Root MSE	service_rate Mean
0.404358	13.14361	0.114646	0.872259

SAS procedure-Booth number

Dependent Variable: service_rate service rate

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	17.3804360	2.4829194	133.06	<.0001
Error	5206	97.1451786	0.0186602		
Corrected Total	5213	114.5256146			

R-Square	Coeff Var	Root MSE	service_rate Mean
0.151760	15.66076	0.136602	0.872259

SAS procedure-Vehicles in Queue

Dependent Variable: service_rate service rate

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	10.8649035	1.8108173	90.96	<.0001
Error	5207	103.6607111	0.0199080		
Corrected Total	5213	114.5256146			

R-Square	Coeff Var	Root MSE	service_rate Mean
0.094869	16.17587	0.141096	0.872259

Figure 9 **R-square for simple regressions with $Y =$ service rates, and $X_1 =$ day, $X_2 =$ time, $X_3 =$ active booth number, $X_4 =$ vehicles in queue, respectively.**