Project Summary – CATSA

Wait Time Impact Model at Pre-Board Screening Checkpoints for Canadian Airports

by Yiqiang Zhao, Patrick Boily, Wenzhe Ye
Centre for Quantitative Analysis and Decision Support, Carleton University

CLIENT ORGANIZATION

The Canadian Air Transport Security Authority (CATSA) ensures the safety and well-being of passengers as they board their flights at Canadian airports each year. A federal crown corporation founded in the aftermath of the September 11, 2001 terrorist attacks on American soil, CATSA protects the public by efficiently screening air travellers and their baggage at designated airports.

CATSA offers a number of security-related services including pre-board screening, where passengers and their belongings are searched for prohibited and potentially dangerous items such as knives, firearms and explosives. CATSA also conducts hold-baggage screening – where checked baggage is screened using explosive detection equipment – and non-passenger screening where airport workers who have access to restricted areas are searched. Finally, CATSA also administers and maintains the Restricted Area Identity Card (RAIC) program.

PROJECT INTENT, SCOPE, AND OBJECTIVES

Numerous factors influence the wait time at pre-board screening checkpoints at Canadian airports: the schedule intensity of departing flights, the volume of passengers on these flights, the number of servers and processing rates at a given checkpoint, etc.

One of CATSA's goals is to ensure that the pre-board screening experience at Canadian airports is made as efficient as possible by minimizing the waiting time at checkpoints. In order to help CATSA gain a better understanding of the waiting process, CQADS developed a Wait Time Impact (WTIM) model using a Queueing Theory approach.

The original scope of the project consisted of:

1. Provide estimates of the passenger arrival rates λ, the processing rates μ and the number of servers c at each checkpoints, using the available field data
2. Calculate the Quality of Service (QoS) level (p_x, x) and determine what service level can be achieved at each checkpoint (i.e. the percentage p of passengers which will wait less than x minutes, for x fixed) for a given arrival rate λ, processing rate μ and number of servers c.
3. Provide the average number of servers c^* required to achieve a prescribed QoS level (p_x, x), given an arrival profile λ^*.
4. Implement the WTIM on a SAS platform to allow for the analysis of various scenarios (such as passenger growth, for instance) via the tweaking of a small number of parameters and whenever the available data is updated.

Upon satisfactory completion of these objectives, a second phase was initiated at CATSA's behest, with the intent of extending their implementation of the WTIM. This second phase's scope consisted of three objectives:

5. Provide quality of service (QoS) level curves $(p_x(x), x)$ (i.e. cumulative distribution curves) under various arrival rate and number of active servers for each checkpoint (where x is allowed to vary).
6. Seamlessly integrate the WTIM with CATSA's scheduling optimizer, in order to implement a one-click SAS program.
7. Provide validation and modeling analytics support for the integrated WTIM.

METHODOLOGY

In order to complete the assignment, CQADS used the following methodological steps:

1. *Exploration of available data*, in order to identify any underlying patterns and essential characteristics.
2. *Understanding the conceptual model*: including document review pertaining to CATSA's existing framework to gain a full understanding of the structure of its queueing system.
3. *Estimation of model parameters*, which required: making appropriate assumptions to simulate the processes in the queueing system according to the knowledge gained through data exploration; selecting appropriate parameter estimation methods, using the appropriate statistical inference and/or numerical method, based on the completeness and characteristics of the existing data; and conducting parameter estimation accordingly.
4. *Implementation of the conceptual $M/M/c$ model* on a SAS platform, which allowed for the discovery of the importance of certain notions whose importance only emerged after running some early scenarios through the modification of a small number of parameters (arrival profile, service time distribution, number of servers, service level, etc.), in particular when it came to vacation policy regarding the number of lines, which lead to a switch to a generalized $M/M/1$ model.
5. *Validation of the generalized $M/M/1$ model*, by comparing the estimated characteristics of the prototype queueing model (e.g. inter-arrival and service time distributions, average idle time per server, etc.) with their empirical counterparts to determine the validity of the conceptual model. The conceptual model was found to be mostly invalid until a key link between the average arrival rate, the processing rate and the number of lines was established. This combined generalized $M/M/1$ and Regression model produced good results in most cases, but in certain instances, a departure from the empirical data could still be identified. Further analysis lead to a breakthrough and the introduction of a Departure parameter. The final model, then, combined the $M/M/1$, Regression and Departure hypotheses.
6. *Performance evaluation of the final model* was achieved in two ways. A preliminary performance evaluation pitted the model favourably against historical data, but the ultimate test came once predictions were compared to data that were collected after the final model was delivered, again very favourably.
7. *Documentation of the final model*: a technical report providing an overview of the model, as well as describing and justifying the various assumptions, was written and delivered to CATSA stakeholders.
8. *Knowledge transfer* was achieved through meetings (in person or by phone) and email exchanges detailing the progress, increasing in frequency as the deadline approached.
9. *Provision of on-going support* to CATSA's model users allowed for a number of improvements, both in scope and in implementation.

PROJECT SUMMARY

The available data covered 26 checkpoints, at 8 Canadian airports. At each checkpoint, the pre-board screening process is structurally similar: passengers arriving at the beginning of the main queue may have their boarding passes scanned at the S_1 position (the start of the waiting queue), but they are always scanned at the S_2 position (as they are being processed).

For each checkpoint, 3 datasets were available for each year:

- the *Raw Data* which contains – for each passenger reaching the end of the queue at S_2 – the date, scan time at S_1, scan time at S_2 and the wait time between S_1 and S_2;
- the *Checkpoint Utilization Report* which records – for each day of the year and each non-overlapping 15-minute block – the maximum number of open processing lines, and

- the *Waiting Time Report* which consists of the subset of the *Raw Data* for which S_1 and S_2 are both available (and for which observations with anomalous and/or outlying wait time behaviour have been removed by CATSA).

The data was then grouped into meaningful clusters exhibiting properties that can be characterized by the same Poisson process, which allows for proper estimation of queueing model parameters, under the assumption that the queueing model $M/M/c$ model was valid.

[One difficulty with this approach is that, in practice, the number of servers c varies with time, according to a vacation policy which depends on a variety of factors. As such, it is extremely difficult to model. This is problematic since the sought QoS level (p_x, x) depends not only on the arrival rates, but also on the processing rates, which themselves depend, among other things, on the number of open servers. Switching to a generalized server (behind which the actual servers are hidden) circumvents this issue, but at the cost of not immediately being able to retrieve the number of servers c from the generalized $M/M/1$ model.]

The average arrival rates λ for each cluster were computed from the *Raw Data* using Burke's Theorem and were shown to indeed follow a Poisson process as the inter-arrival times between consecutive S_2 events were i.i.d. exponential random variables with parameters λ, lending support to the generalized $M/M/1$ hypothesis. The average wait times $\overline{W}_q$ were then estimated using the *Wait Time Report*.

[An analysis of the reasons for the omission of those observations without an S_1 scan from the *Wait Time Report* suggests that using the latter to estimate the cluster average wait times $\overline{W}_q$ is likely to affect the predicted QoS levels, especially in the small wait time regime. However, since short wait times are not likely to cause consternation among the general public, this issue may not arise in practice and can be side-stepped in the estimation phase.]

The estimated processing rates $\hat{\mu}_M$ and QoS levels $(\hat{p}_M, x)$ were easily recovered from the relationships

$$\overline{W}_q = \frac{\hat{\rho}_M}{\hat{\mu}_M - \lambda}, \quad \hat{p}_M = 1 - \hat{\rho}_M e^{-(\hat{\mu}_M - \lambda)x},$$

where $\hat{\rho}_M = \lambda/\hat{\mu}_M$ represents the estimated traffic intensity.

Since these relations do not hold if the generalized $M/M/1$ hypothesis fails, the need to validate it became more pressing. The simplest way to do so was to compare the wait times generated by the model to those of the empirical data: were the estimated QoS curves $\hat{p}_M(x)$ "close to" the empirical QoS curves $p(x)$? Using two different metrics (largest relative difference ratio, largest area ratio), we showed that the generalized $M/M/1$ assumption, while not exact, is a reasonable one to make at the checkpoint level.

[This result is achieved without explicitly invoking the number of open servers c. Granted, that number is implicitly involved in the determination of the average wait times $\overline{W}_q$, but it does not change the fact that it cannot be recovered using solely the tools provided by queueing theory. The Regression assumption asserts that, on a quarterly level, the cluster processing rates $\mu = \mu(c, \lambda)$ is a function of the number of active servers c (hidden behind the generalized server) and the arrival rates λ, and that this functional relationship is the same for all regression clusters making-up a given quarter.]

Using the *Checkpoint Utilization Report*, the average service rates per line $\hat{\mu}_M/c$ and average arrival rates per line λ/c were estimated for each checkpoint, quarter, and cluster, and then regressed against one another to determine the optimal regression parameters $\hat{a}, \hat{b}$, yielding new estimates $\hat{\mu}_R = \hat{a}c + \hat{b}\lambda$ for the cluster processing rates. Thus, estimates for the QoS level $(\hat{p}_R, x)$ were easily computed, without explicitly referring to processing rates, using

$$\hat{p}_R = 1 - \frac{\lambda}{\hat{a}c + \hat{b}\lambda} e^{-(\hat{a}c+\hat{b}\lambda-\lambda)x},$$

which held as a direct consequence of the combined $M/M/1$ and Regression assumptions.

Using the two validation metrics introduced above, it was shown that the combined assumptions, while proving slightly less valid than the $M/M/1$ hypothesis on its own, still provided reasonably close QoS estimates at the quarter and checkpoint levels.

[This lessened validity should come as no surprise, as there is no way to extract the number of clusters c without postulating an external relationship of the form $\mu = \mu(c, \lambda)$, and that the simple linear regression form used necessarily introduces some uncertainty. Some of that uncertainty might decrease using a more complex regression function.]

In order to predict the number of servers required to meet a given QoS level (p, x) at a given checkpoint during a given quarter (i.e. for a given pair of regression parameters a, b), for a given arrival profile λ, it then sufficed to solve for c, yielding

$$c_R = \frac{1}{ax}\left[\mathrm{W}_0\left(\frac{\lambda x}{1-p} e^{\lambda x}\right) - b\lambda x\right],$$

where W_0 is the main branch of the Lambert W-function.

[Unfortunately, W_0 cannot be evaluated by elementary needs except at special values and so one has to depend on efficient numerical algorithms to recover c_R. As SAS does not lend itself particularly well to repeated algorithmic computations, it becomes imperative to find a quick and relatively accurate alternative approach, such as the following approximation, which can be implemented in SAS:

$$c_R \approx \frac{1}{ax}\left[\ln\left(\frac{\lambda x}{1-p} e^{\lambda x}\right) + 1.031 \ln\left(\ln\left(\frac{\lambda x}{1-p} e^{\lambda x}\right)\right) + 0.207 - b\lambda x\right],$$

valid for $e \le \frac{\lambda x}{1-p} e^{\lambda x} \le e^{1000}$.]

For any given checkpoint, quarter, and cluster, it was thus possible to compare the actual number of open servers c (given by the *Checkpoint Utilization Report*), and the estimated value c_R given the actual arrival rate λ and the actual QoS level (p, x). Plotting c_R against c for all clusters strongly suggested that the prediction and the actual values were linked at the checkpoint level according to $c = \hat{d} \cdot c_R$, for some checkpoint departure parameter $\hat{d}$. Computed values of d near 1 for nearly all checkpoints further validated the combined model. The final prediction for the number of servers was further refined by setting $c_D = \hat{d} \cdot c_R$.

In theory, it is thus possible to forecast the number of servers c_D for a cluster using only its regression parameters a, b, its departure parameter d, an arrival rate λ, and a QoS level (p, x). The validation procedure in this case is slightly different: it makes little sense to compare the predicted value c_D with the actual number of servers c found in the historical data as the prediction depends not only on the forecasted arrival rate (which is likely to be different from the historical rate), but also on the attained QoS level (for which an independent forecast is unavailable).

The best validation alternative, then, is to wait for new data to be collected, to determine the actual cluster arrival rate and QoS level to be used in the forecast in order to provide a prediction c_D, and to compare it with the actual c recorded over the data collection period.

DIFFICULTIES AND ADDITIONAL TASKS

Apart from the usual issues surrounding the transfer of large datasets, the specific issue of the lack of algorithmic computability of the Lambert W-function in SAS, and the technically difficult modeling problem, additional tasks were requested by CATSA and the scope of the project was accordingly extended.

These tasks included:

- the modification of the originally requested $M/M/c$ queueing model to a $M/M/1$ queueing model;
- the introduction of the combined $M/M/1$ and regression hypothesis to recover the number of servers c;
- the identification and clean-up of data integrity issues for three airports (YUL, YYC, YVR), and
- the conversion of a MATLAB non-linear solver for the computation of the Lambert W-function to an approximation which could be implemented in SAS.

RESULTS AND RELEVANCE

While CQADS has not been made privy to all details of the validation work conducted by CATSA on 2013 data (especially when it comes to exact accuracy figures), the CATSA Project Authority has let it be known informally that the predictions and QoS level curves which were generated by the WTIM were found to be quite in agreement with the actual data: it is CQADS' understanding that the model is currently in use within Operations Reporting and Analysis at CATSA.

PROJECT LOGISITCS

Timeline

Phase I	Contractual Tasks	14-Jun-14 to 31-Aug-14	
	Additional Tasks (at CATSA's request)	01-Sep-14 to 30-Sep-14	
Phase II	Contractual Tasks	11-Oct-14 To 11-Nov-14	

Resources/Personnel

Yiqiang Q. Zhao, Ph.D.
Professor, School of Mathematics and Statistics
Carleton University
Subject Matter Expert (Queueing Theory)

Patrick Boily, Ph.D.
Managing Consultant, CQADS
Project Lead / Senior Analyst

Wenzhe Ye
Consultant, CQADS
Analyst (Queueing Theory)

Jun Gao
Junior Consultant, CQADS
Analyst

Total Effort Level 580 hours (estimate)

		Zhao	Boily	Ye	Gao
Phase I	Contractual Tasks	60	100	120	50
	Additional Tasks (at CATSA's request)		100		
Phase II	Contractual Tasks		150		
	Total:	**60**	**350**	**120**	**50**

Dollar Value

Phase I	Contractual Tasks	$22,000.00
	Additional Tasks (at CATSA's request)	$22,000.00
Phase II	Contractual Tasks	$10.000.00
	Total:	$54,000.00 (+ HST)

CATSA Project Authority

Maryam Haghighi, Ph.D.
Senior Advisor

Operations Reporting and Analysis
99 Bank Street
Ottawa, Ontario K1P 6B9
Canada

613 993-7094
maryam.haghighi@catsa.gc.ca

Extract from the report

Wait Time Impact Model at Pre-Board Screening Checkpoints for Canadian Airports

by Yiqiang Zhao, Patrick Boily, Wenzhe Ye

[…]

1.2 Model Outline

The model establishes a relationship between the arrival rates, the service rates, the number of servers and the service levels. Basic concepts, process descriptions, and limitations are provided in §2.

The WTIM is best described via the flow chart of Figure 1 on the next page (the various concepts will be defined as they arise in the corresponding section):

1. computation of the arrival rates λ from the raw data (§3.2);
2. computation of the distribution of the number of servers c from the checkpoint utilization reports (§3.3);
3. computation of the waiting time distribution W_q from the waiting time report (§3.4);
4. computation of the QoS levels (p, x) from the waiting time report (§3.4);
5. computation of the estimated QoS levels $(\hat{p}_M, x)$ under the $M/M/1$ assumption (§3.5);
6. validation of the $M/M/1$ assumption based on a comparison of $(\hat{p}_M, x)$ and (p, x) (§3.6);
7. computation of the estimated service rates $\hat{\mu}_M$ under the $M/M/1$ assumption (§3.5);
8. computation of the seasonal checkpoint regression parameters a, b under the combined $M/M/1$ and $Regression$ assumptions (§4.1);
9. computation of the estimated QoS levels $(\hat{p}_R, x)$ under the combined $M/M/1$ and $Regression$ assumptions (§4.2);
10. validation of the combined $M/M/1$ and $Regression$ assumptions based on a comparison of $(\hat{p}_R, x)$, $(\hat{p}_M, x)$ and (p, x) (§4.3);
11. prediction of the number of servers c_R under the combined $M/M/1$ and $Regression$ assumptions (§5);
12. validation of the combined $M/M/1$ and $Regression$ assumptions based on a comparison of c_R and c (§5.3);
13. computation of the checkpoint departure parameters d under the combined $M/M/1$, $Regression$ and $Departure$ assumptions (§5.3);
14. computation of the estimated QoS levels $(\hat{p}_D, x)$ for various projected arrival growth rates λ^* under the combined $M/M/1$, $Regression$ and $Departure$ assumptions (§6.2);
15. prediction of the number of servers c_D for various projected arrival growth rates λ^* under the combined $M/M/1$, $Regression$ and $Departure$ assumptions (§6);
16. final validation of the combined $M/M/1$, $Regression$ and $Departure$ assumptions based on a comparison of $(\hat{p}_D, x)$ and c_D with empirical data (§6.3).

In order to illustrate the WTIM process, the details are worked out on a step-by-step basis for the Domestic/International Checkpoint at the Edmonton International Airport (YEG), based on 2012 data. The results are shown at the end of each section. A summary of results for all checkpoints is also provided, as well as recommendations and suggested next steps.

[…]

2.1 Definitions

The various mathematical concepts to which the report will refer are described below:

- An $M/M/c$ **queueing model** describes a system where arrivals form a single queue and are governed by a Poisson process (the first M), units arriving are processed by c servers and service times are exponentially distributed (the second M).
- A **Poisson process** is a stochastic process where the time between any two consecutive event has an exponential distribution with parameter λ.
- The **arrival rate** is the rate at which passengers arrive for PBS (i.e. passengers per minute), the **service rate** is the processing rate at a screening line (i.e. maximal potential throughput), the **number of servers** is the number of screening lines and the **service level** is the percentage of people waiting less than a given number of minutes at a checkpoint.

2.2 Description of the PBS Process

At each checkpoint, the PBS process is structurally similar: passengers arriving at the beginning of the main queue may have their boarding passes scanned at the S_1 position, but they are always scanned at the S_2 position (see Figure 2).

[…]

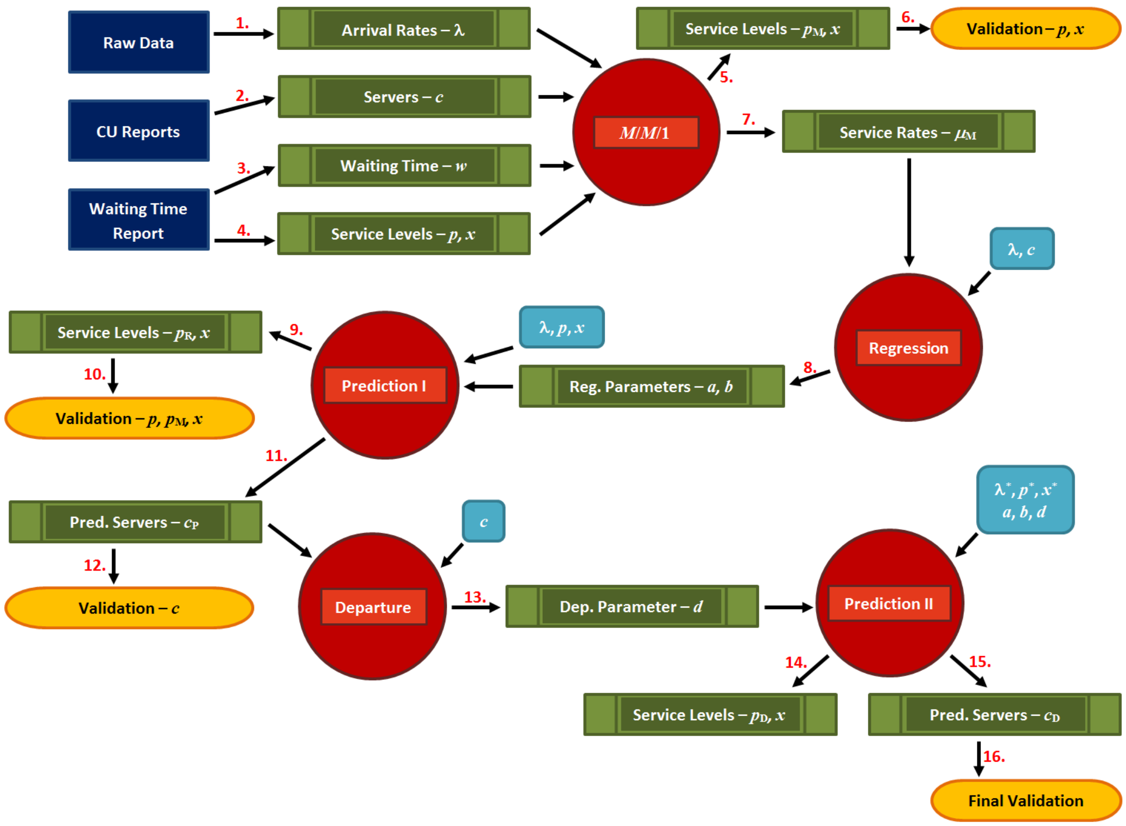

Figure 1 – **WTIM flow.** The dark blue rectangles are CATSA-provided data inputs; the green boxes indicate computed and derived values; the red circles are conceptual nodes; the light blue boxes represent carry-over values, and the orange cells are validation steps.

3. $M/M/1$ Queueing Model

One of the difficulties for the situation under consideration is that the number of servers varies with time, according to different factors: there are times when all servers are busy, others when a number of open servers are idle, and the number of open servers changes according to some vacation policy which it is difficult to model. This is problematic when using an $M/M/c$ model as service rate estimates depend, amongst other things, on the number of open servers.

It is possible to circumvent this issue altogether, without invoking Vacation Models, by noticing that an $M/M/c$ queueing system may be viewed as an $M/M/1$ queueing system where the servers are hidden behind a generalized server (see Figure 3, on the next page). Under that interpretation, the service rates can be estimated independently of the number of servers. Furthermore, not only do $M/M/c$ results still hold for $M/M/1$ (simply by setting $c = 1$ in the appropriate theorems), but the quantities to be computed tend to be simpler in this case.

While this conceptual simplification has removed some of the difficulties associated to server vacation, there remains another problem: the theory of $M/M/1$ systems, alone, is not sufficient to recover (and later predict) the actual (and hidden) number of servers for the checkpoint. This situation can be addressed by finding another way to link the arrival rates, the estimated service rates and the number of servers (see §4.1 for more details).

3.1 Clustering

In order to better predict the average behaviour of a system and its possible outcomes, a wide range of typical patterns must be considered. When analyzing the behaviour of queues, it may become necessary to group the data into meaningful clusters exhibiting similar properties (for example, properties that can be characterized by the same Poisson process).

This approach allows for proper estimation of queuing model parameters (arrival rates, processing rates, etc.), which in turn yields the most reliable results. The selection of the appropriate cluster size relies on finding a balancing point between two extremes:

- In order to properly define the stochastic process, a minimum amount of data with similar properties is required. If clustering is not performed (i.e., if the clusters are too large), the data may present different characteristics which cannot be represented by a single Poisson process.
- On the other hand, if the clusters are too small, they may not contain enough data to capture the underlying properties. More importantly, clusters that cover too short a period are unlikely to exhibit the statistical behaviour of the process.

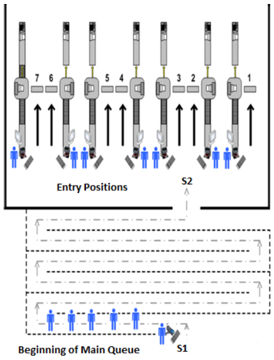

Figure 2 – **Schematics of pre-board screening (PBS).** Passengers enter the main queue, where their boarding pass may be screened at S_1. Once they reach the end of the main queue, their boarding pass is screened at S_2 and they are sent to one of the active lines for processing (image provided by CATSA). In practice, it may happen that only the S_2 reading is available.

A preliminary analysis of the model's accuracy was assessed based on the following criteria: *Checkpoint*, *Weekly patterns* (day of week vs weekday/weekend), *Seasonal patterns* (season vs month) and *Daily patterns* (2-hour period vs 4-hour period). The cluster combination that produced the most encouraging queueing results when compared against actual reports was: checkpoint, weekday/weekend, season, 4 hour-period. Clustering also plays a role in the Regression stage of the model (see §4 for details), but the optimal regression cluster combination need not be the same as the queueing cluster combination.

3.2 Computing the Average Arrival Rate

Since not all boarding passes are scanned at S_1, the Wait Time report (S_1 data) cannot be used to derive the cluster arrival rates. The $S_1 - S_2$ line-up (main queue) is a birth-death process (i.e. a reversible one-dimensional Markov chain). In particular, the forward chain $S_1 - S_2$ and its reverse are stochastically identical and the arrival epochs of the reversed chain are the departure epochs of the forward chain. We can then use Burke's Theorem for $M/M/c$ queues.

Theorem 1 (BURKE'S THEOREM, [1]) *Consider an $M/M/c$ queue in the steady state with arrivals modeled by a homogeneous Poisson process with rate parameter λ. Then the departure process is also a homogeneous Poisson process with rate parameter λ.*

This does not rule out the possibility that, at a particular time, the arrivals at S_1 could be greater than the departures at S_2, due to the inherent randomness of Poisson processes. But all S_1 arrivals will eventually leave at S_2 and thus the fluctuations at S_2 follow the same statistical property governing arrivals to the queue. Therefore, the arrival rates can be estimated by using data readings at S_2 within a given cluster.

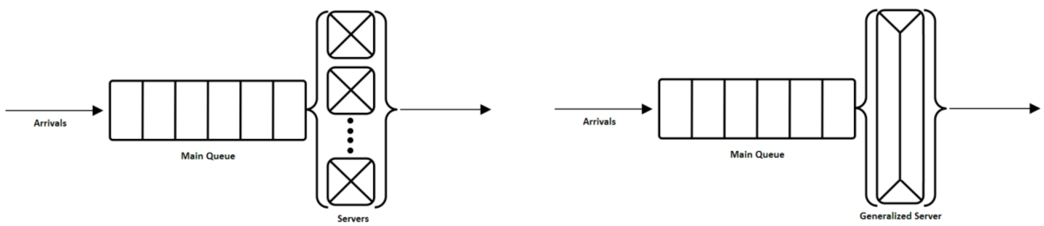

Figure 3 – Queuing systems. Conceptual visualization of an $M/M/c$ queue (on the left) as an $M/M/\mathbf{1}$ queue (on the right); the c servers can be considered as 1 generalized server.

It remains only to show that arrivals follow a homogeneous Poisson process in each cluster (this is a common hypothesis). To do so, one must show, assuming that the number of arrivals in the cluster by time t is denoted by $N(t)$, that (see [3, 4] for details)

1. $N(t)$ is a counting process with independent and stationary increments, and
2. the number of arrivals in any time interval of length t is Poisson-distributed with mean λt, i.e. for all $s, t \geq 0$,

$$P[N(t+s) - N(s) = n] = e^{-\lambda t}\frac{(\lambda t)^n}{n!}, \quad n = 0,1,\ldots$$

The first assumption is satisfied with the introduction of clusters. The third assumption holds if the **inter-arrival times** (the times between consecutive events) are independent and identically distributed (i.i.d.) exponential random variables with the same rate λ: analysis of S_2 in the raw data suggests that this is the case.

The total counts of arrivals for each cluster at YEG's D/I checkpoint based on 2012 data are shown in Table 1. Note that the arrival rate λ is simply calculated by dividing the count in each cluster by the number of minutes in each cluster, independently of the open status of the checkpoint during the period spanned by the cluster. A low arrival rate may thus indicate either that checkpoint traffic was low or intermittent for the cluster, or that it was closed for some or all of the period that it spans.

[…]

Cluster				# of Hours	Count	Avg Arrival Rate
Jan 01 to Mar 31 - YEG (DI) - 2012	Week day	0:00	4:00	260	844	0.055
		4:00	8:00	260	129,069	8.274
		8:00	12:00	260	97,949	6.279
		12:00	16:00	260	84,548	5.420
		16:00	20:00	260	78,964	5.062
		20:00	0:00	260	33,061	2.119
	Week-end	0:00	4:00	104	1,076	0.172
		4:00	8:00	104	39,674	6.358
		8:00	12:00	104	31,200	5.000
		12:00	16:00	104	26,136	4.188
		16:00	20:00	104	28,129	4.508
		20:00	0:00	104	10,013	1.605

Cluster				# of Hours	Count	Avg Arrival Rate
Apr 01 to Jun 30 - YEG (DI) - 2012	Week day	0:00	4:00	260	1,068	0.070
		4:00	8:00	260	128,655	8.247
		8:00	12:00	260	106,704	6.840
		12:00	16:00	260	87,208	5.590
		16:00	20:00	260	82,198	5.269
		20:00	0:00	260	34,330	2.201
	Week-end	0:00	4:00	104	626	0.100
		4:00	8:00	104	35,923	5.757
		8:00	12:00	104	35,683	5.718
		12:00	16:00	104	25,564	4.097
		16:00	20:00	104	24,489	3.925
		20:00	0:00	104	11,735	1.881

Cluster				# of Hours	Count	Avg Arrival Rate
Jul 01 to Sep 30 - YEG (DI) - 2012	Week day	0:00	4:00	260	4,256	0.281
		4:00	8:00	260	128,186	8.345
		8:00	12:00	260	113,577	7.394
		12:00	16:00	260	87,439	5.605
		16:00	20:00	260	82,053	5.260
		20:00	0:00	260	44,213	2.834
	Week-end	0:00	4:00	108	1,781	0.285
		4:00	8:00	108	40,218	6.206
		8:00	12:00	108	41,898	6.466
		12:00	16:00	108	30,237	4.666
		16:00	20:00	108	26,675	4.117
		20:00	0:00	108	15,665	2.417

Cluster				# of Hours	Count	Avg Arrival Rate
Oct 01 to Dec 31 - YEG (DI) - 2012	Week day	0:00	4:00	260	1,114	0.074
		4:00	8:00	260	132,094	8.468
		8:00	12:00	260	102,019	6.540
		12:00	16:00	260	87,806	5.629
		16:00	20:00	260	83,881	5.377
		20:00	0:00	260	35,769	2.293
	Week-end	0:00	4:00	104	771	0.134
		4:00	8:00	104	38,196	6.121
		8:00	12:00	104	38,538	6.176
		12:00	16:00	104	26,683	4.276
		16:00	20:00	104	25,399	4.070
		20:00	0:00	104	11,879	1.904

Table 1 – YEG DI 2012 totals. Number of hours, count of arrivals and average arrival rates, per cluster, per quarter (1st – teal, 2nd – green, 3rd – yellow, 4th – red).

5.3 Validating the Combined Model Using Departure at the Checkpoint Level

The relative accuracy of the formula

$$c_R \approx \frac{1}{ax}\left[\ln\left(\frac{\lambda x}{1-p}e^{\lambda x}\right) + 1.031\ln\left(\ln\left(\frac{\lambda x}{1-p}e^{\lambda x}\right)\right) + 0.207 - b\lambda x\right], \quad \text{when } e \leq \frac{\lambda x}{1-p}e^{\lambda x} \leq e^{1000},$$

used to estimate the average number of servers c_R required to reach the QoS level (p, x) at a checkpoint with regression parameters a, b and average arrival rate λ, suggests another method to validate the combined model.

For any given checkpoint, the plot of c_R against the actual c strongly suggests that the variables are linked according to $c_R = d \cdot c$, for some d.

Linear regressions once again determine the optimal $\hat{d}$ for each checkpoint. The **departure parameter** $\hat{d}$, then, serves as a measure of the predictive model's departure from reality. If $\hat{d} \approx 1$ (i.e. if $c_R \approx c$), then the assumptions that go into the combined model are justified *a posteriori*, in the context of predicting the average number of active servers. The modified predictions $c_D = c_R/\hat{d}$ for a checkpoint where $\hat{d}$ is large or close to 0 (i.e c_R is a poor approximation of c) may still end up being accurate (i.e $c_D \approx c$), but in that case a careful analysis should be undertaken to understand whether any anomalous activity is in play.

The regression of c_R against c for YEG's D/I checkpoint (based on 2012 data) is shown in Figure 6 (on the next page). The departure parameter for this checkpoint is $\hat{d} \approx 1.0161$, which is reflected by the tight linear fit of the two variables. As such, the original prediction c_R requires only a very slight modification.

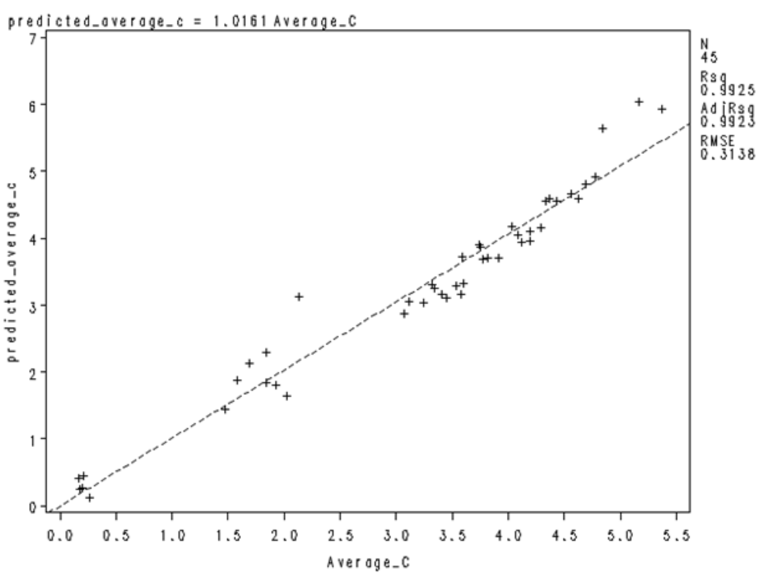

Figure 6 – Departure Regression. Regression of the predicted average number of servers against the actual number of servers.

[…]

8. Recommendations and Final Comments

Perhaps the foremost conclusion is that the $M/M/1$ model on its own provides the best QoS levels estimates, while the best estimates for the average number of active servers are provided by the Departure model.

This discrepancy may be partly explained by the fact that, in any modeling endeavour, some loss of information is inevitable due to the necessity of making simplification assumptions. Below is a list of possible issues which could affect the WTIM's accuracy:

1. The underlying arrival processes are roughly Poisson, and the wait time distributions are roughly conditionally exponential for each cluster; depending on the distance between the theoretical process and the empirical data, the $M/M/1$ assumption may be inappropriate.
2. The wait time distribution may be seriously biased as not every boarding pass has been scanned at S_1, and there is no easy way to verify how representative the subset of those for which wait time data is available actually is.
3. The server vacation policy is unknown, and may not be uniformly adhered to (if one even exists).
4. The actual number of active servers is only crudely approximated by the maximum number of active lines within a 15 minute block.

5. The service rate seems to depend on factors other than the number of active servers and the arrival rate, leading to wildly different outputs for similar inputs and contributing to the lessened accuracy of the regression model when estimating QoS levels.
6. Different checkpoints might require different optimal clustering strategies.

It might be possible to minimize some of that information loss simply by selecting a slightly more sophisticated regression functional form linking the average arrival rate per line and the average service rate per line. Preliminary analysis suggests that the choice

$$\mu = \mu(c, \lambda) = ac + fc^2 + b\lambda$$

may provide better QoS results. Further analysis is needed in that regard, as it is clear that other factors need to be included in order to get the best possible fit and to minimize the number of clusters which become unstable as a result.

Finally, it is conceivable that while adding more historical data to the model could have a useful effect, going too far back into the past may bias the results if policy changes have led to characteristically distinct underlying data over the years. It seems clear that at least one year's worth of data is needed, but, as the datasets only contained trustworthy data for the year 2012, it is still too early to get a definitive answer on this topic.

[…]

References

[3] Newell, G.F. [1971], *Applications of Queuing Theory*, Chapman and Hall.

[4] Ross, S.M. [2010], *Introduction to Probability Models*, 10th ed., Academic Press.

(End of extract)