12-2016

Predicting the Performance of Queues: A Data Analytic Approach

Kum Khiong YANG
Singapore Management University, kkyang@smu.edu.sg

Cayirli TUGBA
Ozyegin University

Mei Wan LOW
Singapore Management University, joycelow@smu.edu.sg

Published in Computers and Operations Research, 2016 Dec, Volume 76, pp. 33-42.
http://doi.org/10.1016/j.cor.2016.06.005

Predicting the Performance of Queues – A Data Analytic Approach

Kum Khiong Yang[1*], Joyce M. W. Low[1], Tugba Cayirli[2]

[1]Lee Kong Chian School of Business Singapore Management University
[2]School of Economics and Administrative Sciences Ozyegin University

kkyang@smu.edu.sg
joycelow@smu.edu.sg
tugba.cayirli@ozyegin.edu.tr

[*]Corresponding author. Lee Kong Chian School of Business, Singapore Management University, 50 Stamford Road, Singapore 17889. Tel: 65-68280369

ABSTRACT

Existing models of multi-server queues with system transience and non-standard assumptions are either too complex or restricted in their assumptions to be used broadly in practice. This paper proposes using data analytics, combining computer simulation to generate the data and an advanced non-linear regression technique called the Alternating Conditional Expectation (ACE) to construct a set of easy-to-use equations to predict the performance of queues with a scheduled start and end time. Our results show that the equations can accurately predict the queue performance as a function of the number of servers, mean arrival load, session length and service time variability. To further facilitate its use in practice, the equations are developed into an open-source online tool accessible at http://singlequeuesystemstool.com/. The proposed procedure of data analytics can be used to model other more complex systems.

Keywords: Data Analytics for Queues, Simulation, Nonlinear regression, Alternating Conditional Expectation.

1. INTRODUCTION

Waiting in line is a common phenomenon in daily life. Customers in banks, postal offices and supermarkets wait in line for services. Cars queue up for gas re-fills and jobs wait for idle machines to start production. Long waiting time leads to unhappy customers and congestion in production facilities. Waiting line or queue management is hence a critical part of both service and manufacturing firms alike. In general, queue management involves a trade-off between the cost of

waiting and the cost of additional service capacity (Jacobs and Chase 2010; Krajewski et al. 2012). The former is often measured by the mean number of customers or mean waiting time in queue while the latter is measured by the mean server utilization or mean overtime per customer served. The probability of a customer being served immediately on arrival is also an indicator of customer service and the proportion of time that at least one server is idle. Management needs to work on a strategy that reduces waiting times and delights its customers without excessive capacity cost.

Queueing theory is a long-standing discipline that is used to derive formulae to compute performance measures, such as mean queue length and customer waiting time under various queueing configurations that can be represented by the Kendall's notation (Hillier and Hillier 2002; Hillier and Lieberman 2004). In its simplest form, the M/M/C model represents a queueing system with C servers and an infinite queue capacity such no customer is rejected. The mean customer arrival rate is characterized as stationary and does not vary with time. In addition, both customer inter-arrival times and service times are assumed to follow exponential distributions, and customers wait in a single queue and are served in the order of first come-first served. The sophistication of such formulae increases if the inter-arrival and service times follow other distributions, such as a General (G) or Erlang (E) distribution, in M/G/C, G/M/C, M/E/C, E/M/C, and E/E/C queueing models.

Nevertheless, many real systems operate in environments where the core assumptions of the above queueing models are violated. Many systems experience nonstationary customer arrival rates that vary both within and across days. For examples, a post office may encounter higher arrival rates during lunch hours and on Mondays. Similarly, an Accident and Emergency Department may receive more patients on Sundays when other clinics are closed. Service systems may also offer services with different service time variability. A specialist clinic, for example, may encounter

higher variable service times than a family clinic. Last but not least, many service systems do not operate continuously, but encounter opening and closing transience every day. Under such circumstances, Yang et al. (2014) show that the formulae derived for the standard M/M/C queueing models fail to provide satisfactory estimates of the system performance.

Whilst our literature review identifies a number of methods that can model the effects of system transience and non-standard queueing assumptions, these procedures are highly complex. The need for an easy-to-use procedure thus remains. This research proposes combining computer simulation and a non-linear regression technique called Alternating Conditional Expectation (ACE) as a procedure for analyzing complex queueing systems. As an illustration, we apply our procedure on a multi-server queueing system that operates with non-exponential service times and with opening and closing transience. Computer simulation is used to generate the data while ACE is used on the data to construct a set of easy-to-use analytical equations to predict the system performance. Once the number of servers, arrival load, session length, and service time variability for such systems are specified, the ACE equations can predict the key performance measures, such as the mean queue length, probability of no waiting on arrival and mean overtime per customer served. Our results show that our proposed procedure is highly accurate in predicting the performance of the queueing systems that are tested. To facilitate the use of the ACE equations, an online open source tool http://singlequeuesystemstool.com/ is developed for practitioners who can use it to estimate the performance of queuing systems that violate the assumptions of exponential service times and continuous operations. The proposed procedure of combining computer simulation and ACE can be used for modeling other more complex systems.

The rest of the paper is organized in the following manner: The next section briefly reviews the related literature on existing methods for predicting the performance of multi-server queues

with system transience. These methods are generally complex with assumptions, which limit their use and accuracy in practice. Section 3 introduces Alternating Conditional Expectation (ACE) as an advanced nonlinear regression technique and develops a set of regression equations for predicting the key performance measures of multi-server systems in the presence of opening and closing transience. Section 4 presents the results on the accuracy of the ACE equations, while Section 5 ends with the conclusions, limitations and suggestions for future research.

2. LITERATURE REVIEW

Queueing theory dates back to the works by Erlang in the early 1900s, and its theoretical development has grown substantially since 1950s with the advances in Operations Research (Hall 1991). The earliest models assume a stationary **M/M/C** queue[1] that operates continuously and are popular for their simplicity and ease of use. They are proposed for a wide range of applications such as telecommunications, manufacturing, and services. Readers can refer to Lakshmi and Iyer (2013) for a comprehensive taxonomy of queueing applications in healthcare services and to Gans et al. (2003) for applications in call centers.

As the literature on queueing is extremely broad, we have to limit our review to works related to our current study. Therefore, we first review literature on the **steady-state** M/G/C **queues and** G/G/C queues, with stationary (i.e. constant) mean arrival and service rates that do not vary with time. In the second part of our review, we discuss the added difficulties of analyzing transient queues with opening and closing of the system, and nonstationary arrival and/or service process.

2.1 Steady-State Queues with Stationary Arrival and Service Process

[1] Note that for consistency, we use a common notation "*C*" for multiple servers, regardless of the original notations used in the discussed papers which may include "*n*", "*m*" or "*s*".

Beyond the stationary **M/M/C** queues, several papers have highlighted the challenges in analyzing non-Markovian queues with non-exponential inter-arrival and/or service times. The emphasis is on finding approximate solutions **given that such generalized models** are analytically intractable (Xiong and Altiok 2009). **The earlier works on the M/G/C queues include Takahashi (1977), which provides approximate formulae for the first two moments of waiting times.** Hokstad (1978) proposes an approximation based on Laplace transform for the steady-state queue length distribution. Several approximations in both light and heavy traffics are introduced and tested by Boxma et al. (1979), **Tijms et al. (1981), and** Kimura (1983, 1996). Recognizing that most of the earlier approximations fail when both the number of servers (C) and the coefficient of variation (CV) are large, especially when traffic intensity is low, Ma and Mark (1995) present a computationally efficient approximation for the mean queue length of M/G/C queues. In their simulation study, Mandelbaum and Schwartz (2002) show that results on M/G/C queues are highly sensitive to the service time distribution.

The more generalized **G/G/C** queues with general inter-arrival and service time distributions have also been studied extensively. Approximations are the most common methods, apart from studies that apply exact numerical methods to phase-type distributions (Takahashi and Takami 1976; Seelen 1986) or **hyper-exponential and Erlang distributions** (v**an Horn and Seelen 1986**). **Kimura (1986, 1994, 2003) provides a number of** approximations **for the** G/G/C queues with infinite and finite queue capacity. It is reported **that these approximations become less accurate when the variability of the inter-arrival and service times increases and when the traffic intensity is low or moderate.** Whitt (1992, 1993) uses an infinite-server approximation and notes that the quality of approximation varies across different performance measures.

A group of related papers addresses the QED (Quality and Efficiency Driven) regime for a special class of queues with high server-utilization and short waiting times relative to service times.

This regime, also known as the Halfin-Whitt regime, is introduced by Halfin and Whitt (1981) in the analysis of a G/M/C queue. Several extensions are developed by Puhalskii and Reiman (2000), Whitt (2004, 2005), Jelenkovic et al. (2004), Mandelbaum and Momcilovic (2008), and Reed (2009). This is a well-studied problem in the context of call-centres, where relevant trade-offs exist between servers' utilization (i.e. efficiency) and customers' waiting times (i.e. quality) (see Gans et al. 2003).

2.2 Mean State of Transient Queues

In Section 2.1, it is concluded that no easy-to-use formulae exist for computing the steady-state performance of both the M/G/C and G/G/C queues. Analysis of **transient queues with the presence of opening and closing of the system and/or nonstationary arrival and service process is even more difficult, as discussed in the following sections.**

2.2.1 Queues with Opening and Closing of the System

In systems that do not operate continuously, the steady-state models may not apply. For example, a clinic may accept patients only from 9 am to 5 pm such that each day begins and ends with an empty queue. Several queueing papers offer transient solutions on how various measures, such as queue length and probability of no waiting, can vary over time. Such transient solutions can then be used to compute the time-average or mean state of the system. While there are many studies on the single-server transient queues, such as Abate and Whitt (1994), Bertsimas and Nakazato (1992), Gong and Hu (1992), Lee and Roth (1993) and Wang (1999), few studies exist on the multi-server transient queues. Kelton and Law (1985) and Murray and Kelton (1988) use an embedded discrete-time Markov chain to calculate the probability of transient queue-length of M/Ph/C queues with exponential inter-arrival times and phase-type service times. Chaudhry and

Zhao (1999) provide an analysis of a transient queue with finite queue capacity. Whitt (2000) analyzes the steady-state and transient performance of an M/G/C queue with a heavy-tailed service time distribution, and shows that the waiting time distribution is also heavy-tailed. Using backward equations of Markov skeleton processes, Hou et al. (2003) study the queue-length distribution of a transient G/G/C queue.

2.2.2 Queues with Non-Stationary Arrival and/or Service Process

It is common in certain systems to observe nonstationary or time-varying behaviour where the mean arrival and/or service rates vary with time. Ibrahim and Whitt (2011) highlight that time-varying arrival rate can introduce significant errors in the estimation of queue length and waiting time when the system experiences alternating periods of overload and underload. A pioneering work in this area is by Jagerman (1975), which estimates the blocking probabilities using a non-stationary Erlang loss model by substituting the expected number of busy servers at specific time points obtained from a non-stationary infinite server model for the system load in a corresponding stationary finite server model. Several other approaches are developed for approximating the performance of non-stationary M/M/C queues. Some of the well-known methods include the Simple Stationary Approximation (SSA), Pointwise Stationary Approximation (PSA), Stationary Independent Period by Period (SIPP), lagged PSA and lagged SIPP (see Green and Kolesar 1991, 1995, 1997; Green et al. 2001, 2003; Whitt 1991; Thompson 1993). Ingolfsson et al (2007) provide a survey and comparison of several exact and approximation methods for the non-stationary M/M/C systems. More recent studies by Jimenez and Koole (2004) and Stolletz (2008) address temporarily overloaded systems in which earlier models do not offer accurate results. Margolius (2005) derives an integral equation for the transient probabilities and mean queue-length of a non-stationary M/M/C queue. Several studies also exist for queues with Erlang service times (Escobar et al. 2002), general inter-arrival and/or service times (Jennings et al. 1996; Green et al. 2007). **Overall, the**

procedures put forth for computing the performance of queue with system transience are complex even for professionals who are trained in Mathematics and Operation Research due to their complicated and recursive computation requirement.

Based on our literature review in **Sections 2.1 and 2.2, we conclude that** existing methods for predicting the performance of queues with non-standard inter-arrival and service times are too complex for widespread applications, especially under the presence of system transience. The complexity of these methods provides the motivation to develop an alternative procedure. To the best of our knowledge, no past research has used the method of data analytics of combining computer simulation and non-linear regression to model such queues. To illustrate the use of our procedure, we develop a set of easy-to-use equations that can accurately predict the performance of a multi-server queue with opening and closing transience. The same procedure can be used to develop equations for queues in other scenarios.

3. RESEARCH METHODOLOGY

Alternating Conditional Expectation (ACE) is an advanced nonlinear regression technique developed by Breiman and Friedman (1985) who won the 1985 best JASA Theory and Methods paper award for their work. The ACE algorithm is designed to produce the best-fitting additive model by estimating an individual smooth transformation for each variable in a regression model to maximize the correlation between the dependent and independent variables. Unlike other empirical methods (Box and Cox 1964; Tukey 1982), ACE transformations are unambiguously defined and estimated without the use of heuristics, restrictive distribution assumptions, or restriction of transformation of a particular parametric family. The reader is referred to Sum et al. (1995) for further details on the superiority of ACE over the standard Ordinary Linear Regression (OLR) technique.

To develop the ACE equations to estimate the mean queue length, probability of no waiting on arrival, and mean overtime per customer served, data must first be collected on the performance of queueing systems under different scenarios. Instead of collecting the data from real systems, computer simulation offers a reliable and convenient platform for collecting the data. The data collected from simulation models may in fact represent the real systems more accurately without the sporadic idiosyncrasies that can occur in some real systems. With the data collected from the simulation models, the equations for predicting the performance of queueing systems can then be developed using the non-linear regression technique ACE.

3.1 Data Generation

A simulation model of a multi-server system with a single queue is built using the simulation software ARENA (Kelton et al. 2010). The objective is to create a generic model that mimics the behavior of real systems that operate with opening and closing of the system. While a number of empirical studies have shown that service times are lognormal with different coefficients of variation, there is no similar evidence on the arrival times (Mandelbaum and Schwartz 2002; Cayirli and Veral 2003). Inter-arrival times of customers are thus still often assumed to be exponential given the dearth of empirical evidence to suggest otherwise (Law and Kelton 2000). Therefore, in this study, we model inter-arrival times as exponential and service times as lognormal. Our model can however be extended fairly easily to include non-exponential inter-arrival times and/or systems with non-stationary mean arrival and service rates.

Following the results of a previous study (Yang et al. 2014), four independent factors are chosen to represent systems with opening and closing of the system: (i) number of servers, (ii) mean arrival load, (iii) target session length, and (iv) service time variability (i.e., coefficient of variation). To consider their impact on the system performance, these factors are modeled at

different values. The number of identical servers is examined from 1 to 10 servers. The mean arrival load is examined from 60 to 100% of the total servers' capacity to represent systems that operate at different utilization. The target session length is examined from 10μ to 50μ time units per session. Once this target session length is reached, new arrivals are denied entry into the system but the actual session ends only after the last customer in the system is served. With no loss of generalizability to other μ, the service times of customers are assumed with a mean of $\mu = 1$ and a coefficient of variation (CV) of 0 to 0.8.

The simulation model is run to collect performance measures of the multi-server system for 750 environments that are represented by combinations of the four independent variables, namely the number of servers in the system (NS), mean arrival load (AL), target session length (SL), and coefficient of variation of service times (CV), as shown in Table 1. For each environment, the simulation model is run for 20,000 clinical sessions to collect the mean performance measures within 1% of the point estimates at 95% confidence level. The three collected performance measures, i.e. dependent variables, include the mean queue length[2], probability of no waiting on arrival[3], and mean overtime per customer served[4].

3.2 ACE Equations

An implementation of the ACE algorithm is available in DBANK, a data management software written for Microsoft Windows, and can be downloaded at http://www.tsDbank.com. In theory,

[2] The mean queue length measures the average number of customers waiting in queue that a customer will encounter on entering the system. The associated mean queue time of customers (QT) can be computed using the Little's law.

[3] Customers can begin service immediately without waiting if they arrive at the system with at least one idle server. The probability of entering service immediately on arrival is a measure of the proportion of time when there is at least one idle server in the system.

[4] Mean overtime per customer served is computed as the amount of overtime measured in server-minutes per customer served. It represents the variable overtime cost that a firm will incur in the form of servers' overtime pay to operate the system beyond its target closing time.

ACE cannot fit a model that is worse than Ordinary Linear Regression. If the variables are related linearly, ACE will simply suggest linear transformations, i.e. no transformations, for the variables. The ACE algorithm in DBank is used to develop the equations for predicting the performance of the queueing system as a function of the four independent variables, namely the number of servers (NS), mean arrival load (AL), target session length (SL), and coefficient of variation of service times (CV).

To develop the nonlinear regression models between the performance measures and the independent variables, we first prepare a set of determinant variables consisting of the four independent variables and their first-order products and divisions. We then use ACE and a build-in stepwise variable selection procedure in DBANK to select the determinant variables that produce the best fit for each performance measure. DBank is run to read in the data, execute the ACE algorithm, and subsequently produce graphical transformations of each performance measure and its selected determinant variables. The approach used is a forward-backward stepwise inclusion and deletion of determinant variables whereby each determinant variable is included or deleted one by one in ranked order of its individual variable adjusted R^2 until no substantial improvement in adjusted R^2 is recorded. This approach also aids to prevent over-fitting the regression models.

As ACE suggests the models in graphical form, mathematical functions have to be suggested to transform the original data to replicate the graphical transformations. Figure 1 shows the graphical transformations suggested by ACE to maximize the model fit between the mean queue length (QL) and the selected determinant variables, i.e. NS, NSxAL, ALxSL, SL/AL and CV/AL. Using the graphical transformations suggested by ACE in Figure 1, we suggest mathematical functions to transform the original data and then use Ordinary Least-Squares on the transformed data to arrive at the following ACE equation (1) for mean queue length (QL) with an adjusted R^2 of 0.997:

$$QL = e^{-2.31 - 0.2627\,NS^{1.3} + 0.00332\,(NS \times AL)^{1.1} + 0.2071\,(AL \times SL/\mu)^{0.4} - 4.5\,(SL/[AL \times \mu])^{0.6} + 77\,(CV/AL)^{1.3}} \quad (1)$$

If the mean queue length is regressed against the same selected variables without the transformations suggested by ACE, OLR will produce an adjusted R^2 of only 0.7759. Using the transformations suggested by ACE thus improves the model fit significantly from an R^2 of 0.7759 to 0.9970. As in all regression models, the residuals between the observed and fitted values are plotted and checked before the model is said to fit the data well. The residuals are plotted and checked for (i) non-linearity of the regression function, (ii) non-constancy of error variance, (iii) presence of outliers, (iv) non-independence of residuals, (v) non-normality of residuals, (vi) omission of independent variables, and (vii) extra independent variables. The inclusion of extra variables in the final models is controlled at least partially by the forward-backward stepwise variable selection procedure used to build the models, which mitigates the risk of multi-collinearity.

To predict the probability of no waiting on arrival (PN) and mean overtime per customer served (OT), two other regression equations are developed. The graphical transformations suggested by ACE for the two performance measures and their selected determinant variables are shown in Figures 2 and 3. Using the same approach as above for developing the equation for mean queue length (QL), the models for probability of no waiting on arrival (PN) and mean overtime per customer served (OT) are developed with an adjusted R^2 of 0.9982 and 0.9908, respectively, in equations (2) and (3):

$$PN = -10.8 + 108/\{1 + 0.8e^{-2[1.38 - 0.00138\,|NS \times AL - 96| - 0.088(AL \times SL/\mu)^{0.36} + 1.29\,(SL/[AL \times \mu])^{0.54} - 1.44e^{-CV \times CV \times \mu/SL} + 10\,(AL/NS)^{-0.79}]}\} \quad (2)$$

$$OT = \mu \times e^{\{-1.05 + 0.0012\,[NS^4 - 26.63\,NS^3 + 253.9\,NS^2 - 938.5\,NS] - 0.283\,SL/\mu - 0.00000116\,[0.00001\,(NS \times AL)^4 - 0.02616\,(NS \times AL)^3 + 24.36\,(NS \times AL)^2 - 8400\,(NS \times AL)] + 0.2\,(NS \times CV \times CV)^{0.69} + 0.0019\,(ALxSL/\mu) + 0.075\,(SL/[NS \times \mu])^{0.59} + 7.2\,(SL/[AL \times \mu])^{1.34} + 240\,|CV/AL - 0.0013|^{1.48}\}} \quad (3)$$

Since OT is a ratio of the difference between the actual and target session lengths divided by the mean number of arrivals per server during the target session length, we can extend the use of equation (3) to estimate the mean actual session length (AS):

$$AS = SL + (AL \times SL)/(100 \times \mu) \times OT \qquad (4)$$

4. RESULTS

In Section 4.1, we analyze the accuracy of the proposed ACE equations (1) to (4) in modeling the performance of M/G/C queues in presence of opening and closing transience. The analysis is done in two steps. First, we present the goodness of fit results on how well the ACE equations fit the original data set of 750 scenarios summarized in Table 1 (i.e., combinations of 6 NS x 5 AL x 5 SL x 5 CV). Next, we validate the accuracy of the ACE equations in estimating the queue performance of a set of 81 independent cases summarized in Table 2 (i.e., 3 NS x 3 AL x 3 SL x 3 CV). Finally, in Section 4.2, we illustrate using our proposed ACE equations with a numerical example, followed by a sensitivity analysis on the choice of service time distributions.

4.1 Accuracy of the ACE Equations

Table 3 summarizes the goodness of the fit results on how well the ACE equations fit the original data set for the four performance measures, namely the mean queue length, probability of no waiting on arrival, mean overtime per customer served, and mean actual session length. For the mean queue length (QL), the mean absolute percentage error is 8.8% with 57 out of the 750 cases (i.e., 7.6%) reporting absolute percentage errors larger than 20% and a worst case of 37.8%. While this may seem significant, the actual error between the estimated and actual queue length is actually quite small, with a mean absolute error of 0.114 customers and a maximum absolute error of 0.636

customers. A closer examination shows that most of the high absolute percentage errors occur in scenarios with short (i.e. close to zero) mean queue lengths, which exaggerate the size of error as a percentage of a short mean queue length. Equation (1) thus fits the mean queue length (QL) of the original data fairly well for the 750 scenarios tested.

No waiting occurs when the number of customers in the system is fewer than the number of servers such that the arriving customer can be served immediately without waiting or joining the queue. Using equation (2) to fit the probability of no waiting (PN) for the 750 scenarios in Table 1, we find that the mean absolute percentage error is 3.5% and all absolute percentage errors are less than 20% with a worst case of 15.9%. Similarly, cases with high absolute percentage errors occur mainly in scenarios with low probability of no waiting, which exaggerate the size of error as a percentage of a low probability of no waiting. Overall, the equation (2) for PN fits the original data even more accurately than equation (1) for QL. The mean absolute error and maximum absolute error of the probability of no waiting are 1.3% and 5.2%, respectively.

Using equation (3) to fit the mean overtime per customer served (OT) for all scenarios in Table 1, we find that only 11 cases (i.e., 1.5%) have absolute percentage errors larger than 20% with a worst case of 27.4%. Thus Equation (3) for OT accurately fits the mean overtime per customer served in the original data with absolute percentage error of less than 20% for 98.5% of the 750 scenarios. The mean absolute error and maximum absolute error of the mean overtime per customer served (OT) are only 0.0062μ and 0.0405μ, respectively. Equation (4) for AS also fits the actual mean session length very accurately in the original data set with a mean absolute percentage error of only 0.5%. No cases are detected with absolute percentage errors worse than 20% and the maximum absolute percentage error for the worst case is only 2.79%. The mean absolute error and

maximum absolute error of the actual mean session length are as small as 0.132μ and 0.708μ, respectively. In other words, the error is less than one service slot.

Next, we present the validation results in Table 4 using a second set of independent data generated through simulation with parameters summarized in Table 2. For the mean queue length (QL), the mean absolute percentage error is 6.9% and only 2 out of the 81 cases (i.e., 2.5%) report absolute percentage errors larger than 20% with a worst case of 22.1%. The actual absolute errors are very small, with the mean and maximum absolute error equal to 0.134 and 0.376 customers, respectively.

Using equation (2) to predict the probability of no waiting on arrival (PN), there are no cases with absolute percentage errors greater than 20%. The mean and maximum absolute percentage errors are 3.2% and 8.3%, and the mean and maximum absolute errors are 1.4% and 3.9%, respectively. Similarly, equations (3) & (4) provide accurate results with no cases with absolute percentage errors greater than 20% for the mean overtime per customer served (OT) and the actual mean session length (AS). Equation (3) estimates OT with a mean absolute percentage error of 6.2% and a maximum absolute percentage error of 13.9% for the worst case. The mean and maximum absolute errors are 0.0062μ and 0.0405μ, respectively. Equation (4) for AS results in a mean and maximum absolute percentage error of only 0.4% and 1.3%; and a mean and maximum absolute error of 0.073μ and 0.293μ, respectively. Overall, these results confirm that the ACE equations provide accurate estimates for the four performance measures of interest.

In summary, the proposed ACE equations can be used successfully for predicting the performance of M/G/C systems with opening and closing transience in terms of the mean queue length, probability of no waiting on arrival, mean overtime per customer served, and mean actual session length once the system parameters are specified. An online tool is developed for

practitioners and is accessible as an open source at http://singlequeuesystemstool.com/ to facilitate its adoption (See Figure 4). Since the equations are derived through regression, it should be noted that they are only valid for system parameters within the range of factor levels in Table 1.

4.2 A Numerical Example with Sensitivity Analysis on Service Time Distributions

As an illustration, let us assume a system with 5 servers, a mean arrival load of 85 percent, mean service time per customer (μ) of 15 minutes and a coefficient of variation of 0.5, and a target session of 25 service slots (i.e. a target session length of 25μ minutes, i.e. 375 minutes). Substituting NS = 5, AL = 85, μ = 15 min., CV = 0.5, and SL = 375 min. into equations (1) to (4), the mean queue length, probability of no waiting on arrival, mean overtime per customer and actual session length can be calculated as 1.683 customers, 40.86%, 1.293 minutes and 402.482 minutes (i.e. 6.708 hours) respectively. The mean queue time of customers (QT) is computed as 5.941 minutes using Little's law (See Figure 4). These estimates correspond very closely to the actual values observed from the simulation model using a lognormal service time distribution, i.e. QL = 1.693 customers, PN = 42.26%, OT = 1.156 minutes, and AS = 6.673 hours.

Next, we test the sensitivity of our results to the choice of service time distributions. The above analysis is repeated with Gamma and Uniform service time distributions using the same coefficient of variation of 0.5. The performance measures are QL = 1.687 (1.713) customers, PN = 42.20% (42.13), OT = 1.149 (1.117) minutes, and AS = 6.657 (6.646) hours for Gamma (Uniform) service times. These results again correspond very closely with the estimates from the ACE equations (1) to (4). Several additional tests are conducted to test the accuracy of these equations across different service time distributions and a range of different environments within Table 1. The results affirm the findings of Ho and Lau (1992) that system performance is affected primarily by the coefficient of variation of its service time distribution but not by its skewness, kurtosis or other

shape parameters. Overall, we conclude that our proposed equations (1) to (4) remain accurate for different service time distributions within the range of environments tested.

5. CONCLUSIONS

Queueing theory is a well-established methodology for analyzing the performance of queues in various application areas, including telecommunications, manufacturing, and services. Specifically, the basic queueing models, such as the stationary M/M/C models are very popular in both academia and practice given their simplicity in predicting queue performance using closed-form formulae. However, prior research shows that the estimation errors of such queueing models are substantial even when only one of the underlying assumptions is violated (Yang et al. 2014). Although many exact and heuristic methods have been suggested in the literature to capture the effects of system transience and non-standard assumptions, these methods are too complicated and demanding for the average users. Consequently, simulation is still often used for its flexibility in modeling the complexity of real systems when one or more of the underlying assumptions of the basic queueing models are violated. While simulation modeling has become popular in recent years given the increasing power of computers and availability of user-friendly simulation software, building a simulation model is still a relatively time-consuming and daunting endeavor for practitioners who lack the expertise.

The notion of data analytics, combining the advantage of computer simulation with statistical methods to construct easy-to-use formulae for complex systems, is thus highly appealing. With this purpose in mind, this research proposes combining computer simulation and Alternating Conditional Expectation, or ACE in short, as a data analytic tool to construct equations that can estimate the performance of analytically intractable queues, such as a M/G/C queueing system with opening and closing transience. Specifically, once the system parameters, such as the number of

servers, mean arrival load, target session length and coefficient of variation of service times are specified, the practitioners can use the proposed ACE equations to estimate the key performance measures, such as the mean queue length, probability of no waiting on arrival, mean overtime per customer served, and mean actual session length.

Our results show that ACE, which does not force or assume any linear or non-linear relationship among the independent and dependent variables, produces equations that consistently yield higher adjusted R^2 values than Ordinary Linear Regression for all the performance measures considered. Compared with the simulation results using the original plus an independent data set, we show that the ACE equations predict the queue performance very accurately, with small mean and maximum absolute errors. Although lognormal distribution is used to model the service times, sensitivity analysis reveals that our results are robust to the choice of service time distributions. As a result, we conclude that the proposed ACE equations can be used to accurately estimate the performance of multi-server queueing systems with stationary Poisson arrival rates and general service times in presence of opening and closing transience.

Some limitations of the current study include the assumptions made in the simulation model for collecting data for the ACE models. Exponential customer inter-arrival time is a common assumption in the literature and is suggested as fairly representative of the random arrival patterns in many real systems. On the other hand, the assumption of stationary arrival rate is more restrictive given the time-varying arrivals in many real systems. Although such situations are omitted from the current work, these factors can be incorporated fairly easily into the simulation models and ACE equations as independent variables. Future research can also extend the application of data analytics of combining computer simulation and ACE to other more complex queueing systems. Testing our procedure against others in the literature is beyond the scope of our current study, but it is a worthwhile endeavor for future research.

REFERENCES

Abate J, Whitt W (1994) Transient behaviour of the M/G/1 workload process. Operations Research 42(4), 750-764.

Bertsimas DJ, Nakazato D (1992) Transient and busy period analysis of the GI/G/1 queue: the method of stages. Queueing Systems 10(3), 153–184.

Boxma OJ, Cohen JW, Huffels N (1979) Approximations for the mean waiting time in an M/G/s queue system. Operations Research 27, 1115–1127.

Box GEP, Cox DR (1964) An analysis of transformations. Journal of the Royal Statistical Society B26, 211 -252.

Breiman L, Friedman JH (1985) Estimating optimal transformations for multiple regression and correlation (with discussion). Journal of the American Statistical Association 80, 580–619.

Cayirli T, Veral E (2003) Outpatient-scheduling in health care: A review of the literature. Production and Operations Management 12(4), 519-549.

Chaudhry ML, Zhao YQ (1999) Transient solutions of some multi-server queueing systems with finite spaces. International Transactions in Operational Research. 6(2), 161-182.

Escobar M, Odoni AR, Roth E (2002) Approximate solution for multi-server queueing systems with Erlangian service times Computers & Operations Research, 29, 1353-1374.

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

Gong WB, Hu JQ (1992) The Maclaurin series for the GI/G/1 queue. Journal of Applied Probability 29, 176–184.

Green L, Kolesar P (1991) The pointwise stationary approximation for queues with nonstationary arrivals. Management Sciences 37(1), 84–97.

Green L, Kolesar P (1995) On the accuracy of the simple peak hour approximation for Markovian queues. Management Science 41(8), 1353–1370.

Green L, Kolesar P (1997) The lagged PSA for estimating peak congestion in multiserver Markovian queues with periodic arrival rates. Management Science 43(1), 80–87.

Green L, Kolesar P, Soares J (2001) Improving the SIPP approach for staffing service systems that have cyclic demands. Operations Research 49(4), 549–564.

Green L, Kolesar PJ, Soares J (2003) An improved heuristic for staffing telephone call centers with limited operating hours. Production and Operations Management 12(1), 46-61.

Green L, Kolesar P, Whitt W (2007) Coping with time-varying demand when setting staffing requirement for a service system. Production and Operations Management 16(1), 13–39.

Halfin S, Whitt W (1981) Heavy-traffic limits for queues with many exponential servers. Operations Research, 29(3), 567-588.

Hall RW (1991) Queueing Methods for Services and Manufacturing. (1st Edition). Prentice Hall.

Hillier FS, Hillier MS (2002) Introduction to Management Science: A Modeling and Case Studies Approach with Spreadsheets. (2nd Edition). McGraw-Hill/Irwin.

Hillier FS, Lieberman GJ (2004) Introduction to Operations Research. (8th Edition). McGraw-Hill.

Ho CJ, Lau HS (1992) Minimizing total cost in scheduling outpatient appointments. Management Science 38(12), 1750–1764.

Hokstad P (1978) Approximation for the M/G/m queue. Operations Research 26, 511–523.

Hou Z, Yuan C, Zou J, Liu Z, Luo J, Liu G, Shui P (2003) Transient distribution of the length of GI/G/N queueing systems. Stochastic Analysis and Applications 21 (3), 567-592.

Ibrahim R, Whitt W (2011) Wait-time predictors for customer service systems with time-varying demand and capacity. Operations Research 59(5), 1106-1118.

Ingolfsson A, Akhmetshina E, Budge S, Li Y, Wu X (2007) A survey and experimental comparison of service level approximation methods for non-stationary M/M/s queueing systems INFORMS Journal of Computing 19(2), 201–214.
Jacobs FR, Chase RB (2010) Operations and Supply Chain Management. (13th Edition). New York: McGraw-Hill/Irwin.
Jagerman DL (1975) Nonstationary Blocking in Telephone Traffic. Bell System Technical Journal 54(3), 625–661.
Jelenkovic P, Mandelbaum A, Momcilovic P (2004) Heavy traffic limits for queues with many deterministic servers. Queueing Systems 47, 53–69.
Jennings O, Mandelbaum A, Massey W, Whitt W (1996) Server staffing to meet time-varying demand. Management Science 42(10), 1383-1394.
Jimenez T, Koole G (2004) Scaling and comparison of fluid limits of queues applied to call centers with time varying parameters. OR Spectrum 26 (3) 413–422
Kelton WD, Law AM (1985) The transient behavior of the M/M/s queue with applications for steady-state simulation. Operations Research 33(2), 378–395.
Kelton WD, Sadowski RP, Swets NB (2010) Simulation with Arena. (5th Edition), New York: McGraw-Hill.
Kimura T (1983) Diffusion approximation for an M/G/s queue. Operations Research 31(2) 304-319.
Kimura T (1986) A two-moment approximation for the mean waiting time GI/G/s queue. Queue Management Science 32, 751-763.
Kimura T (1994) Approximation for multi-server queues: system interpolations. Queueing Systems 17, 347–382.
Kimura T (1996) A transform-free approximation for the finite capacity M/G/s queue. Operations Research 44(6) 984-988.
Kimura T (2003) A consistent diffusion approximation for finite capacity multi-server queues. Mathematical and Computer Modeling 38, 1313–1324.
Krajewski LJ, Ritzman LP, Malhotra MK (2012) Operations Management – Processes and Supply Chains. (10th Edition). Pearson.
Lakshmi C, Iyer SA (2013) Application of queueing theory in healthcare: A literature review. Operations Research for Healthcare 2(1-2), 25–39.
Law AM, Kelton WD (2000) Simulation Modeling and Analysis. (3rd Edition) New York: McGraw-Hill.
Lee IJ, Roth E (1993) A heuristic for the transient expected queue length of Markovian queueing systems. Operations Research Letter 14(1), 25–27.
Ma BNW, Mark JW (1995) Approximation of the mean queue length of an M/G/c queueing system. Operations Research 43(1) 158-165.
Mandelbaum A, Schwartz R (2002) Simulation experiments with M/G/100 queues in the Halfin-Whitt (QED) regime. Technical Report, The Technion, Haifa, Israel.
Mandelbaum A, Momcilovic P (2008) Queues with many servers: The virtual waiting-time process in the QED regime. Mathematics of Operations Research 33(3), 561-586.
Margolius BH (2005) Transient solution to the time-dependent multiserver Poisson queue. Journal of Applied Probability 42 (3), 766-777.
Murray JR, Kelton WD (1988) The transient behavior of the M/E_k/2 queue and steady state simulation. Computers and Operations Research 15(4), 357–367.
Puhalskii AA, Reiman MI (2000) The multiclass GI/PH/N queue in the Halfin-Whitt regime. Advances in Applied Probability 32(2), 564–595.

Reed J (2009) The G/GI/N queue in the Halfin-Whitt regime. The Annals of Applied Probability 19(6) 2211-2269.
Seelen LP (1986) An algorithm for Ph/Ph/c queues. European Journal of the Operations Research Society 23, 118-127.
Stolletz R (2008) Approximation of the non-stationary M(t)/M(t)/c(t)-queue using stationary queueing models: The stationary backlog-carryover approach. European Journal of the Operations Research Society 190 (2), 478–493.
Sum CC, Yang KK, Ang JSK, Quek SA (1995) An analysis of Material Requirements Planning (MRP) benefits using Alternating Conditional Expectation (ACE). Journal of Operations Management 13(1), 35–58.
Takahashi Y, Takami Y (1976) A numerical method for the steady state probabilities of a GI/G/C queueing system in a general class. Journal of the Operations Research Society of Japan 19(2), 147–157.
Takahashi Y (1977) An approximation formula for the mean waiting time of an M/G/c queue. Journal of the Operations Research Society of Japan 20, 150-163.
Thompson GM (1993) Accounting for the multi-period impact of service when determining employee requirements for labor scheduling. Journal of Operations Management 11(3), 269–287.
Tijms H, van Hoorn M, Federgruen A (1981) Approximations for the steady-state probabilities in the M/G/c queue. Advance Applied Probability 13, 186-206.
Tukey JW (1982) The use of smelting in guiding re-expression. In: Launer, R. L. and Siegel, A. F. (Editors). Modern Data Analysis (pp. 83-102). New York: Academic Press.
van Hoorn MH, Seelen LP (1986) Approximations for the GI/G/c queue. Journal of Applied Probability 23(2), 484-494.
Wang CL (1999) On the transient delays of M/G/1 queues. Journal of Applied Probability 36(3), 882–893.
Whitt W (1991) The pointwise stationary approximation for Mt/Mt/s queues is asymptotically correct as the rates increase Management Science 37, 307-314.
Whitt W (1992) Understanding the efficiency of multi-server service systems. Management Science 38(5), 708–723.
Whitt W (1993) Approximations for the GI/G/m Queue. Production and Operations Management 2(2), 114–161.
Whitt W (2000) The impact of a heavy-tailed service-time distribution upon the M/GI/s waiting-time distribution. Queueing Systems 36(1), 71–87.
Whitt W (2004) A diffusion approximation for the G/GI/n/m queue. Operations Research 52(6), 922–941.
Whitt W (2005) Heavy-traffic limits for the G/H*2/n/m queue. Mathematics of Operation Research 30(1), 1–27.
Xiong W, Altiok T (2009) An approximation for multi-server queues with deterministic reneging times. Annals of Operations Research 172, 143-151.
Yang KK, Low JMW, Cayirli T (2014) Modeling queues with simulation versus M/M/C models. Journal of Service Science Research 6(1), 173–192.

Figure 1: Graphical Transformations suggested by ACE for Mean Queue Length

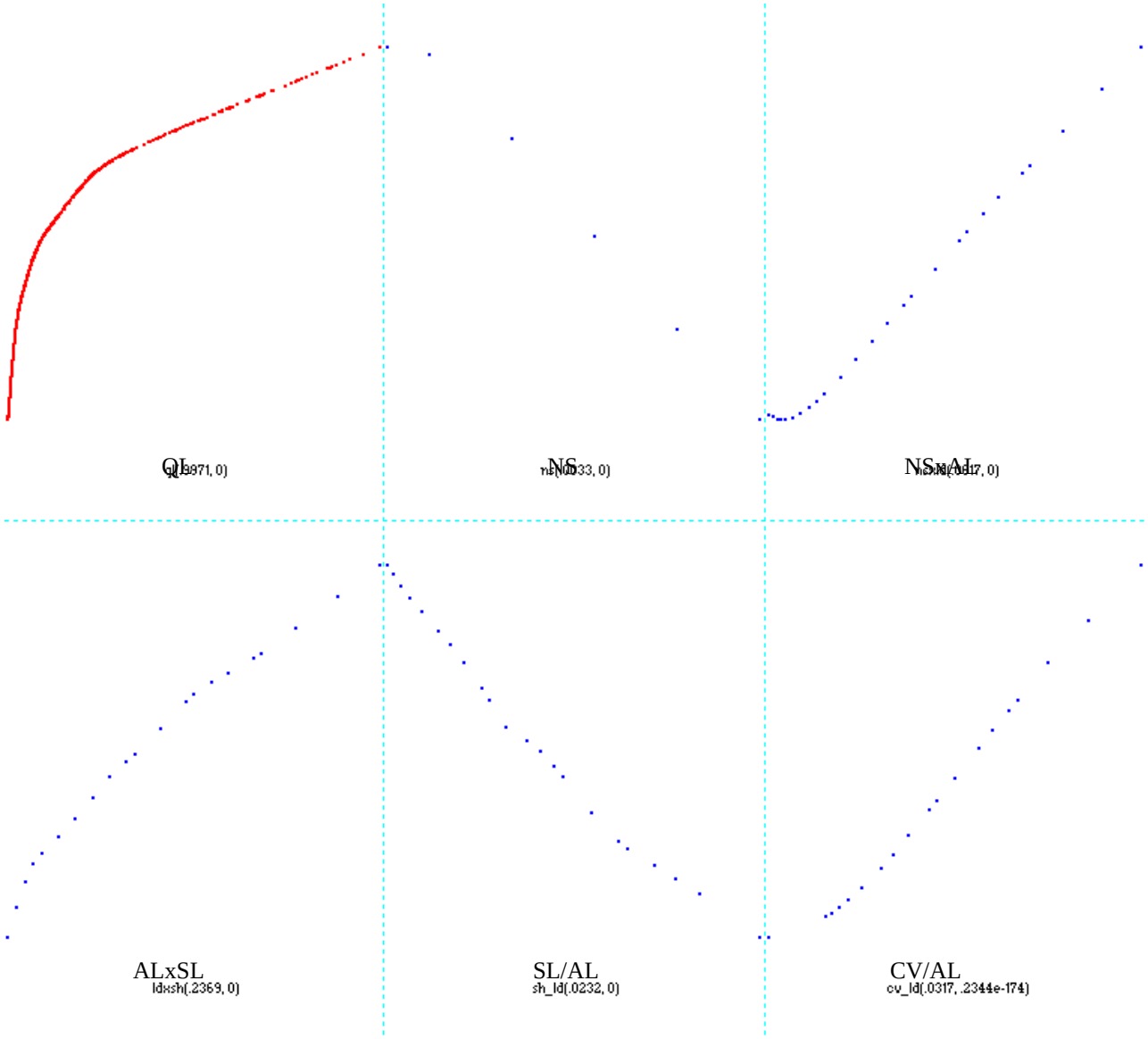

Figure 2: Graphical Transformations suggested by ACE for Probability of No Waiting on Arrival

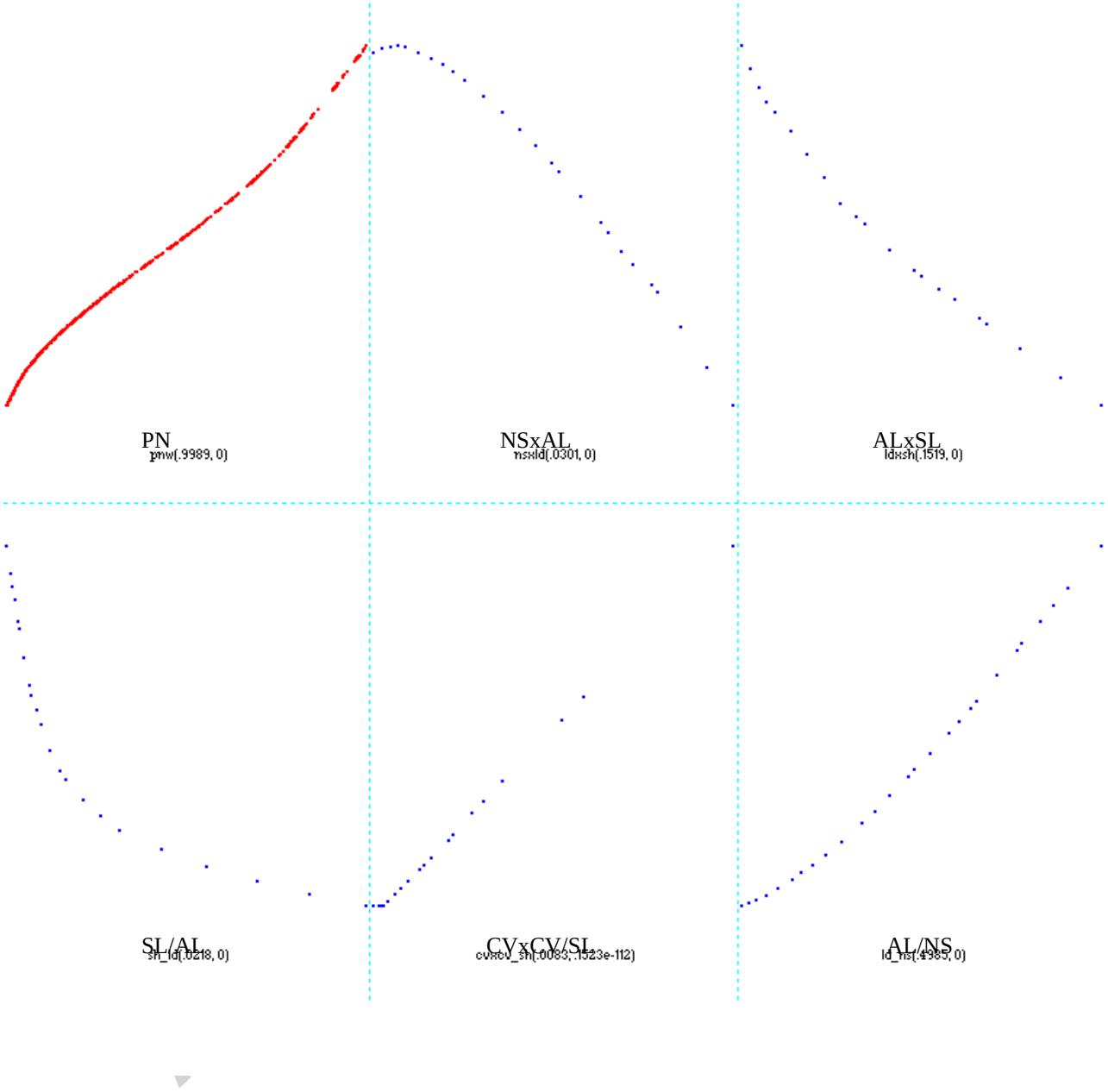

Figure 3: Graphical Transformations suggested by ACE for Mean Overtime per Customer Served

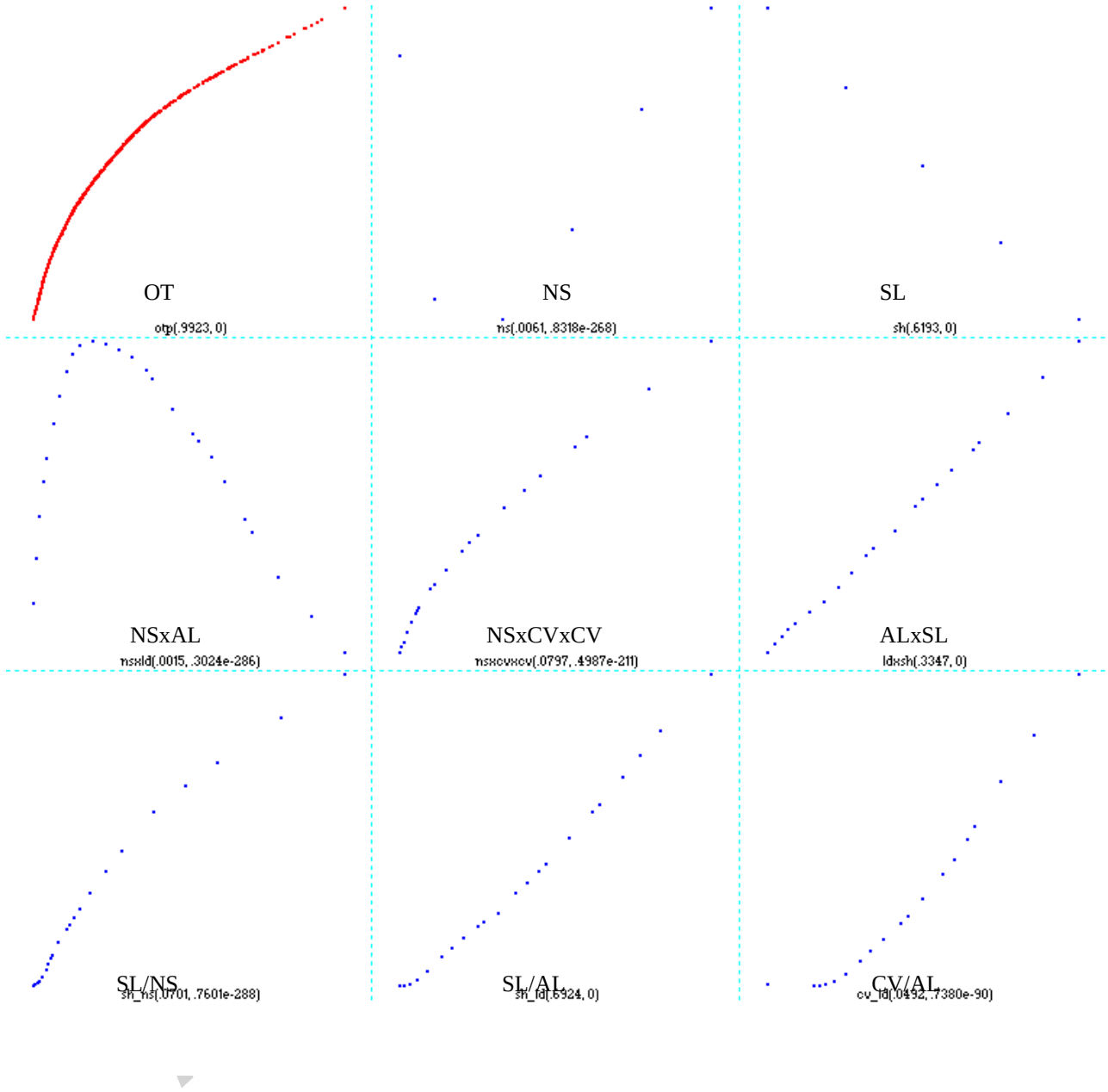

Figure 4: Online Tool for M/G/C Model with Opening and Closing Transience

Number of Servers (NS)	5	
Arrival Load (AL)	85	%
Mean Service Time (μ)	15	min.
Standard Deviation of Service Time (σ)	7.5	min.
Session Length (SL)	375	min.

Press to Compute Need Help?

QUEUE PERFORMANCE		
Mean Queue Length (QL)	1.683	
Mean Queue Time (QT)	5.941	min.
Probability of No Waiting on Arrival (PN)	40.86	%
Mean Overtime per Customer Served (OT)	1.293	server-min.*
Mean Session Overtime (SO)	27.482	min.
Mean Actual Session Length (AS)	402.48	min.

* Server-minute is a measure of server capacity. For example, 10 server-minutes can be 2 servers working for 5 minutes or 1 server working for 10 minutes.

Table 1: Experimental Design

Factors	Levels
Number of Servers (NS)	1, 2, 4, 6, 8 & 10 servers
Arrival Load (AL)	Exponential inter-arrival times with mean load of 60, 70, 80, 90 & 100 (%)
Session Length (SL)	10μ, 20μ, 30μ, 40μ & 50μ
Coefficient of Variation of Service Times (CV)	Lognormal with mean μ = 1 and CV = 0, 0.2, 0.4, 0.6 & 0.8

Table 2: Test Set of 81 Cases

Factors	Levels
Number of Servers (NS)	3, 5 & 7 servers
Arrival Load (AL)	Exponential inter-arrival times with mean load of 75, 85 & 95 (%)
Session Length (SL)	15μ, 25 & 45μ
Coefficient of Variation of Service Times (CV)	Lognormal with mean μ = 1 and CV = 0.3, 0.5 & 0.7

Table 3: Estimation Errors of the ACE Equations Based on Original Data Set[1]

	Mean Absolute Percentage Error	Number (%)[1] of Cases with Absolute Percentage Error > 20%	Maximum Absolute Percentage Error for Worst Case	Mean Absolute Error	Maximum Absolute Error
Mean Queue Length (QL)	8.8%	57 (7.6%)	37.8%	0.114 customers	0.636 customers

Probability of No Waiting on Arrival (PN)	3.5%	0 (0%)	15.9%	1.3%	5.2%
Mean Overtime per Customer Served (OT)	7.2%	11 (1.5%)	27.4%	0.0062μ	0.0405μ
Mean Actual Session Length (AS)	0.5%	0 (0%)	2.79%	0.132μ	0.708μ

[1] Sample size is 750 cases.

Table 4: Estimation Errors of the ACE Equations Based on Independent Data Set[2]

	Mean Absolute Percentage Error	Number (%)[1] of Cases with Absolute Percentage Error > 20%	Maximum Absolute Percentage Error for Worst Case	Mean Absolute Error	Maximum Absolute Error
Mean Queue Length (QL)	6.9%	2 (2.5%)	22.1%	0.134 customers	0.376 customers
Probability of No Waiting on Arrival (PN)	3.2%	0 (0%)	8.3%	1.4%	3.9%
Mean Overtime per Customer Served (OT)	6.2%	0 (0%)	13.9%	0.0055μ	0.0175μ
Mean Actual Session Length (AS)	0.4%	0 (0%)	1.3%	0.073μ	0.293μ

[2] Sample size is 81 cases.

Highlights

- We propose data analytics as a method for analyzing complex systems.
- Proposed method combines simulation to generate the data and a non-linear regression technique to analyze the data.
- Method can be used on multi-server queuing systems with system transience and non-standard service time.
- The estimation equations developed are very accurate and easy to use.
- Proposed method can be used for modeling other complex systems.