DATA-BASED ESTIMATION OF PARAMETERS FOR TIME-INHOMOGENEOUS CTMC BIRTH-AND-DEATH MODELS OF CALL CENTERS

Maciej Rafal Burak
West Pomeranian University of Technology in Szczecin
email: maciej.burak@zut.edu.pl ORCID:0000-0003-0214-8650

This work has been originally published at the European Simulation and Modelling Conference 2023 (ESM'2023) as a conference paper in the Conference Proceedings.
Please cite as: Burak, M.R., 2023. *DATA-BASED ESTIMATION OF PARAMETERS FOR TIME-INHOMOGENEOUS CTMC BIRTH-AND-DEATH MODELS OF CALL CENTERS. 37th Annual European Simulation and Modelling Conference, ESM 2023, Toulouse.* 67-75.

KEYWORDS
uniformization, inhomogeneous CTMC, call center, balking, abandonment

ABSTRACT

In this paper, we present a novel approach to estimate parameters of a queuing model using real call center data. Our approach leverages the transient solution of an inhomogeneous continuous-time Markov chain (CTMC) queuing model emulating the behavior of the genuine system.

Specifically, to assess the fidelity of our modeling approach, we use an example of approximating the real call center system with an inhomogeneous $M_n/M/c/K+M$ CTMC model. To do this, we utilize authentic call center data for replicating the behavior of the real system through our model. In particular, the model incorporates the true time-dependent rates for the arrival process, service rates, and the number of available servers. Our analysis focuses on assessing the accuracy of the Markovian assumptions made for modeling customer abandonment during waiting periods.

Furthermore, we investigate the performance of our model under two distinct scenarios: overloaded systems and systems operating in a quality-driven mode. By examining these cases, we ascertain the effectiveness of our assumptions in accurately representing the behavior of the call center.

Finally, we demonstrate the practical application of our findings by showcasing how a simple and computationally efficient $M_n/M/c/K+M$ Markovian approximation of a real call center can be used for accurate personnel planning while adhering to service-level constraints.

INTRODUCTION

The problem of determining the required number of call center agents to handle an (time-variable but to some extent predictable) amount of service-requests (calls) is the most classic one within queuing theory. It inspired pioneering work of A.K.Erlang (Erlang 1917) which have laid foundation to the theory of queues and is still a subject of numerous academic publications in various disciplines of operational research.

One of the challenges in designing of appropriate queuing models of such systems is how to model customer abandonment. Given the inherent challenges associated with precise modeling of empirical patience distributions, it is customary to resort to approximations. The quality of such approximations is measured based on their ability to estimate performance measures with high accuracy, compared to the outcome of the real system operating under the same input data, i.e., the number of incoming calls, service rate, and number of servers, which can all vary with time and system state. Additionally, desirable features of these approximate queuing models include computational efficiency, as they need to be repeatedly computed during optimization processes to estimate agent schedules, and ease of parameter estimation based on historical data.

The earliest proposal by Palm (1957) usually referred to as (Palm-)Erlang-A model assumed exponential distribution of patience time. It can be represented as $M/M/S+M$ in Kendall's notation. According to different studies of real life call center data the probability distribution of abandonment times is clearly non exponential, moreover it depends strongly of type of service and configuration of the system. Consequently we cannot assume a single patience time distribution. For example Mandelbaum and Zeltyn (2013) presents example of uncensored data caused by a technical problem resulting in all customer calls being unanswered for a whole day and described the resulting shape of patience distribution as idiosyncratic. In Jouini et al. (2013) the authors analyze data of four different call centers with the conclusion that distributions of the patience time are not exponential and specific for different call centers and that exponential distribution severely underestimates the abandonment. In Feigin (2006), Jouini et al. (2011) the authors show that information received while waiting e.g., about projected waiting time or about the possibility to call later influences the waiting behavior of customers.

An important property when modeling call centers for practical purposes is their inherent time-variability.

This includes the volume of incoming calls which is usually modeled as inhomogeneous Poisson process with arrival rates varying depending on time of the day, day of the week etc., which can be to some extent predicted in advance, based on historical data. Based on these forecasts the required number of agents in function of time can be estimated using appropriate models in order to guarantee that particular performance measures (service level) are met. For an overview of common performance measures used in service level agreements (SLA) see for example Jouini et al. (2013). The planning process results finally in an agent schedule i.e., which agents are assigned to which shifts on particular days. An overview of and references for the personnel planing process can be found for example in Aksin et al. (2007), for call arrival forecasting in e.g., recent Petropoulos and et al. (2022). The agent schedule determines only the maximum number of available agents. The true instantaneous number of available agents is usually smaller, as it accounts for other activities requiring the agents to use so called breaks (e.g., data entry, answering e-mails, outgoing calls etc.) or rest time which is usually regulated by labor law (private breaks). This allows to some extent real-time adjustments based on the current situation e.g., allowing more offline activities when there is no customers waiting or asking the agents to postpone their meal breaks when the number of waiting customers or waiting time in the queue exceeds a certain limit value necessary to maintain service level. Often agents are allowed to shorten the time of the call from the optimal level defined in the SLA, in order to increase the number of answered calls if necessary. Despite of the system varying in time, most proposals regarding more accurate modeling of general data based distributions of abandonment assume the system being stationary (e.g.,Jouini et al. (2013) or recent Kanavetas and Balcıoglu (2022)). There are different proposals how to approximate time-variable systems by using stationary models, which are in fact very popular in the literature as shown for example in Defraeye and Nieuwenhuyse (2015). Unfortunately, such stationary approximations often cannot adequately model time-varying systems (see e.g., Green et al. (2001) or Ingolfsson et al. (2010)). Other approaches using inherently time-inhomogeneous queuing models with general distributions including discrete event simulation or numerical solution of CTMCs with phase-type distributions, which can approximate any positive-valued distribution with arbitrary accuracy, can be challenging computationally for bigger models (e.g., Cezik and L'Ecuyer (2008), Creemers et al. (2014)).

In our previous works Burak and Korytkowski (2020) and Burak (2018), we introduced a modified uniformization algorithm that enables highly efficient calculation of transient values for time-inhomogeneous systems, which can be represented using Markovian birth-and-death queuing models.

In the literature, various proposals exist for birth-and-death queuing models approximating abandonment behavior. A simple example is the previously mentioned (Palm-)Erlang-A model, which assumes an exponential distribution for customer patience. Another extension incorporates balking, where customers immediately leave the system without waiting for service. This extension has been described by Deslauriers et al. (2007) and Jouini et al. (2013), among others. In this study, we utilize this model to demonstrate the effectiveness of our parameter fitting method with real data.

Other computationally simple models that have the potential to approximate general distributions are state-dependent Markovian birth-and-death queuing models. For instance, Whitt (2005) proposes a method to derive state-dependent abandonment rates by leveraging the original time-to-abandon distribution of an $M/M/S/K+GI$ queue, enabling the construction of its $M/M/S/K+M(n)$ approximation. A similar approach was recently employed in Kanavetas and Balcıoglu (2022). Another state-dependent approximation of general distributions is the $M(n)/M(n)/S+M(n)$ queue, introduced by Brandt and Brandt (2002). Regrettably, the authors of these models provide parameter fitting methods that assume stationarity of the system.

The objective of this paper is to develop a methodology for fitting parameters in a simple queuing model to ensure accurate modeling of customer abandonment based on real call center data. The resulting approximation should yield precise outcomes for the number of abandoned calls, even in the presence of time-inhomogeneous input data (such as the number of incoming calls, number of servers, and service rate), without assuming system stationarity at any given point in time.

STATISTICAL ANALYSIS OF CALL CENTER DATA

One of the most important aspects in the area of modeling and quantitative assessment of call centers is measuring the quality of delivered service. Defining the desired level of quality is an important part of business agreements, often formalized through a "Service Level Agreement" (SLA). SLAs are particularly significant in the outsourcing industry, as they have financial implications and require ongoing quality assurance inspections to ensure compliance with the contract.

Among the commonly utilized Key Performance Indicators (KPIs), the Service Level (SL) or Telephone Service Factor is by far the most common one. This metric measures the proportion of calls answered within a specified waiting time (e.g., 80/20, indicating that 80% of calls should experience a waiting time of no more than 20 seconds). However, if Service Level were the sole KPI, some call centers might deviate from the First-Come-First-Served (FCFS) principle and prioritize calls with waiting times below the threshold. Unfortunately,

this would inevitably result in prolonged waiting times for customers who have already exceeded the specified threshold or may even lead to call abandonment. To mitigate such issues, SLAs encompass additional KPIs, with one of the most prevalent being the percentage of abandoned calls.

In the literature, authors such as Baccelli and Hebuterne (1981) and Whitt (1999) have made a distinction between two types of customers abandonment in queuing systems. The first group consists of customers who quickly decide not to wait for service, commonly referred to as *balking* customers. The second group comprises customers, who initially choose to wait, but eventually abandon the queue due to a loss of patience.

Regarding Key Performance Indicators (KPIs), some authors, like Jouini et al. (2013), propose excluding balking customers from the calculation of KPIs, as their decision not to wait does not necessarily indicate dissatisfaction. Balking customer do also to some extent improve the Service Level by effectively decreasing the number of incoming calls versus the service rate of the system. On the other hand, the percentage of customers who initially choose to wait but later abandon the queue (known also as *reneging* customers) can be seen as an important indicator of perceived service quality, as their abandonment suggests dissatisfaction. It is worth noting that reneging customers tend to exhibit quite a high level of patience, as observed in empirical data analysis e.g., by Feigin (2006). Therefore, once they exceed a certain threshold, they no longer significantly influence the Service Level.

In this paper, our focus will be on estimating the total number of customers who leave the system without receiving the service, encompassing both cases of abandonment. However, if required by the KPIs specified in the Service Level Agreement (SLA) of a particular organization, it is possible to distinguish and separate balking customers by using an arbitrary time threshold.

In order to demonstrate our method, we utilized real data obtained from an existing call center operated by a bank in the United States. The data was made available in the form of a relational database by the *Technion - Israel Institute of Technology* (*https://seecenter.iem.technion.ac.il/databases/USBank/*). Specifically, we focused on a single customer class or skill known as *Summit*. For detailed information on the data extraction process, please refer to Appendix.

Out of the 420,700 calls routed to the *Summit* skill during this timeframe, 399,278 calls were successfully handled by agents ($service\_time > 1$), while 16,616 calls were abandoned by customers (abandonment rate = 3.95%) and 4,806 calls were prematurely terminated by agents ($service\_time \leqq 1$ *and* $outcome = 2$). Figure 1 illustrates the percentage of served calls based on the waiting time experienced by customers. Notably, we observe that 82% of calls were serviced within a waiting time of 20 seconds, indicating that the target service level for this call center was probably set at 80/20.

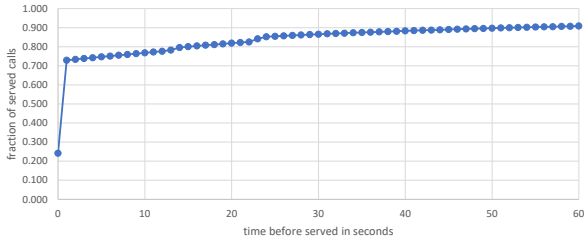

Figure 1: served calls by experienced waiting time

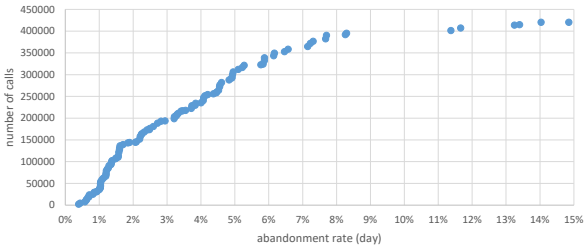

Figure 2: number of calls by daily abandonment rate

We have chosen the day as the fundamental planning period for our analysis. This choice is based on the assumption that the schedule has been planned appropriately, taking into account the expected traffic intensity. Consequently, potential slight deviations in expected traffic intensity during the day, can still be to some extent managed through real-time adjustments of available agents. Further, we assume that the possible real-time adjustments are consistent and known to the agents i.e., in a particular situation like higher than expected call volume, which can negatively impact the service levels, we can expect same reaction e.g., decisions about reducing offline activities (breaks) or regarding the recommended call duration. In Figure 2, we present the distribution of days categorized by their corresponding abandonment rates. We can see that, during the analyzed period, 50% of the calls were handled on days with abandonment rates lower than 3.3%.

DATA BASED PARAMETER FITTING FOR CTMC CALL CENTER MODEL

In order to show our proposal for parameter fitting based on real data, we propose the following approximate model of a call center similar to the proposal in Burak and Korytkowski (2020): the analyzed period is finite (e.g. one working day) with the system starting empty. Customers arrive according to an inhomogeneous Poisson process with rate $\lambda(t)$, the service time is i.i.d. exponentially distributed with time-dependent rate $\mu(t)$. If there are free servers, the customer is served immediately, otherwise customers being put in

the queue can either leave the system (hang up or balk) immediately with probability 1-γ, reducing the arrival rate or, after joining the queue, they abandon (renege) after reaching their *patience time*. Queued requests are served FCFS (first-come-first-served). Served requests rejoin the queue when a server leaves. The state variable $X(t)$ represents the total number of service requests (served/waiting calls) in the system at time t. The size K of the system is finite and equal to $s(t)$ = number of identical servers plus $q(t)$ = the capacity of the queue. The CTMC model we will apply is an $M_t(n)/M_t/S_t/K + M$ queue with state dependent arrival rate equivalent to patience distribution only with discrete probability mass at zero representing customer who experience waiting in the queue enabling them to disconnect before the call is picked up by the agent (balk). If the customer decides to wait, then he either will be served or will abandon with exponentially distributed rate similarly to the Palm/Erlang-A model. Because $X(t) = k$ is a birth-and-death process, it can be described by the following state-dependent birth $q_{k,k+1}(t) = \lambda_k(t)$ and death $q_{k,k-1}(t) = \mu_k(t)$ rates:

$$\lambda_k(t) = \begin{cases} \lambda(t), & \text{if } 0 \le k \le s(t) - 1 \\ \gamma\lambda(t), & \text{if } s(t) \le k \le K - 1 \end{cases} \quad (1)$$

$$\mu_k(t) = \begin{cases} k\mu(t), & \text{if } 1 \le k \le s(t) - 1 \\ s(t)\mu(t) + (k - s(t))\eta, & \text{if } s(t) \le k \le K, \end{cases} \quad (2)$$

resulting in a tridiagonal time-dependent generator matrix $Q(t)$. The transient distribution at time t $\pi(t) = [\pi_0(t), ..., \pi_K(t)]$ representing the probabilities of the system being in any of the states at time t, can be described by the modified Chapmann-Kolmogorow forward equations 3.

$$\frac{d\pi(t)}{dt} = \pi(t)Q(t) \quad (3)$$

The time dependent arrival and service rates and number of available servers are derived from the true call center data as described in Appendix and aggregated by 60s periods. The resulting time-dependent changes in $Q(t)$ are discrete and therefore we can replace the time-inhomogeneous CTMC with a sequence of homogeneous systems computing state probability vectors for consecutive time periods recursively, using uniformization as described in Burak and Korytkowski (2020).

To demonstrate our approach for parameter estimation, we initially select a specific day in which the call center operated in Quality and Efficiency Driven (QED) mode throughout the entire day. The chosen example day is June 13, 2001, with an abandonment rate of 1.04%. The extracted data shows that the number of servers in the call center dynamically adjusts according to the state of the system. Specifically, if there are typically more than 2-3 calls waiting in the queue, agents will change their status from break to available to handle the queued calls and prevent the waiting time from exceeding the service level threshold. Conversely, when there are no calls waiting in the queue, agents who have completed their calls may switch their status to break, allowing them to engage in off-line work. The number of available agents is always lower than the total number of logged-in agents. As a result of these real-time adjustments, customers experienced only short waiting times in the queue. Moreover, there were no idle (waiting) agents, which allowed customers who chose not to use the service (referred to as balking customers) to disconnect before their calls were picked up by an agent or alternatively during the transfer to the agent i.e., the call is disconnected by the customer and service time $<$ 2s. All customers who decided to wait were served i.e., there was no abandonment. Because of the lack of reneging in this case we use a CTMC model with abandonment rate η set to zero reducing the model to a $M(n)/M/S/K$ queue. Figure 3 shows the results of the model - the expected system state calculated as:

$$ESS(t) = \sum_i i\pi_i(t), \ \pi(t) = [\pi_0(t)..\pi_K(t)] \quad (4)$$

is plotted together with the real system state (*state*) and the number of servers *s*. The model was calculated with the system size K=200. The number of abandoned (balking) calls calculated by the model fits exactly the real number for $\gamma = 0.978$. We will further assume this rate as the natural balking rate for all systems where approximately all customers experience waiting.

The second data set that we use to fit the model parameters comes from a day where there was insufficient number of agents planned to handle the incoming calls and perform necessary (e.g. call related) off-line tasks. Specifically we choose the data from June 1, 2001, with abandonment rate of 13.21%. Due to the significant abandonment resulting from unacceptable waiting times in this particular case, modeling the data using the model settings from the first example (QED mode) leads to highly inaccurate results. Specifically, since no abandonment is considered in the model, the calculated system state quickly reaches values close to the system size K. Consequently, the modeled system becomes overloaded and unable to accept new calls, resulting in call rejections or blocks due to a full queue as depicted in Figure 4.

In order to model correctly the observed abandonment resulting from unacceptable waiting time we have to introduce the abandonment process with exponentially distributed rate η. The results of the model for this scenario are depicted in the Figure 5. Assuming the same balking rate as in the first example, we determined that an abandonment rate of $\eta = 0.05$ (corresponding to an expected patience time of 1200 seconds) yields the same number of abandoned calls as observed in the real system. We further tested the same model parameters for

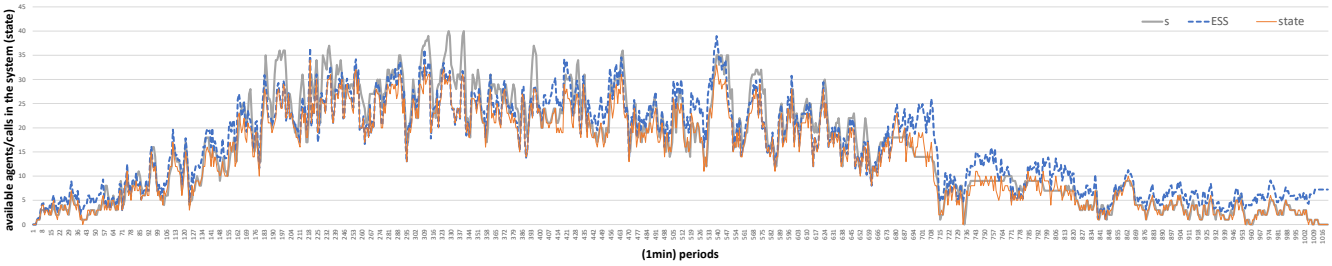

Figure 3: 13 VI $M(n)/M/S/K$

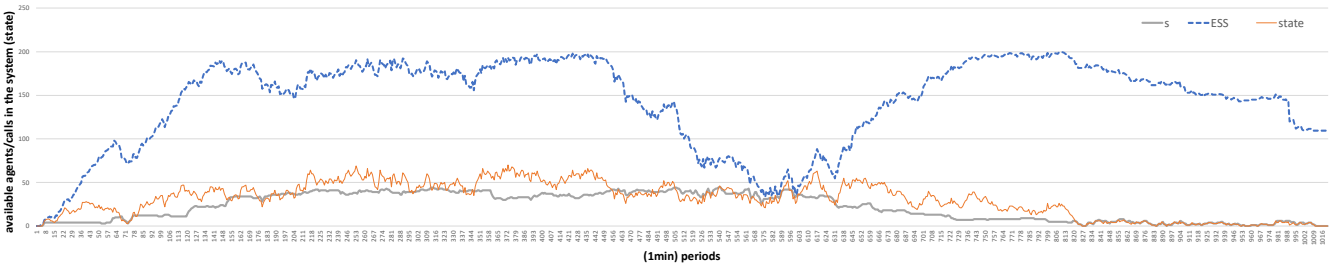

Figure 4: 1 VI $M(n)/M/S/K$

other days with high abandonment rates ranging from 11,5% to 15% with similar results i.e. the estimation error of the abandonment rate compared to the real data was smaller than 3%.

However, applying now the same model parameters again to the QED example would yield highly inaccurate results, with the modeled abandonment rate being three times higher than the actual rate. This discrepancy arises from the difference in the distribution of the abandonment process employed in the model compared to the empirical data. In the model, an exponentially distributed abandonment rate is assumed, with the highest probability of abandonment occurring for short waiting times. In reality, customers who choose to wait after deciding not to balk are less likely to abandon the call until a certain waiting threshold is reached. As a result, the model significantly overestimates the abandonment rate, even for small probabilities of the system becoming overloaded. This well known property of models incorporating abandonment with exponential rate (e.g., Erlang-A) is also the reason for dominance of the simpler Erlang-C model in practical applications.

If we assume the average values of abandonment and service level from historical data as the target levels for planning future time periods, the model parameters should produce accurate results for "typical" days, ensuring compliance with service level constraints, such as SL=80/20 and abandonment rates below 3.95% in our case. This is a common approach when using stationary queuing models, where deviations from average results are attributed to the variability of stochastic input variables representing the arrival, service, and abandonment processes i.e., depending on their average values in function of time and the corresponding probability distribution.

When employing the proposed method of matching real data of a particular "typical" day with a time-inhomogeneous CTMC model to estimate the optimal model parameters, a naive approach would be consequently to choose a day with a "typical" level of abandonment. By examining the distribution of served calls by the percentage of abandoned calls on a particular day, as illustrated in Figure 2, one could select a day close to the median, representing a full day with an abandonment rate lower than that of 50% of the served call volume, approximately 3.3% in our case.

For example, Figure 6 displays real data (number of available agents and number of calls in the system) for two consecutive days, namely April 17, 2001 and April 18, 2001, with abandonment rates closely aligned with the median (2.82% and 3.22% respectively). These days serve as an illustrative choice when estimating the optimal model parameters using the matching time-inhomogeneous CTMC model, aligning the observed data of a "typical" day.

However, upon closer examination of this example, it becomes evident that waiting only occurs during specific time periods when the number of available servers does not match the offered load (i.e., exceeding the possibilities of real-time adjustments). These periods, such as 207 to 244 or 280 to 295 on the 17th, or 2138 to 2207 on the 18th, are clearly not the result of random variations in the arrival process throughout the day. Consequently, contrary to the approach proposed by most authors, it

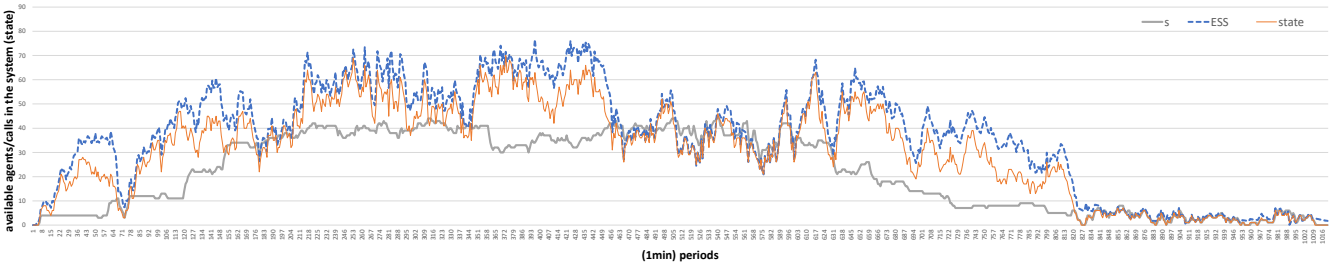

Figure 5: 1 VI $M(n)/M/S/K + M$

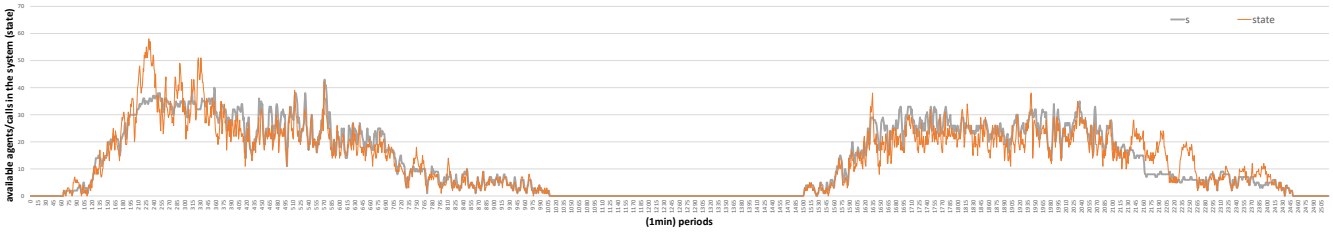

Figure 6: 17-18.IV system state

becomes clear that the service level achieved by a real call center is primarily determined by the ratio of time periods with minimal waiting to time periods with an overloaded state due to an insufficient number of agents (weighted by the number of calls in each respective period), rather than being solely the result of stochastic character of the modeled system.

This conclusion is further supported by the distribution of waiting times experienced by served calls over the entire analyzed period, as presented in Figure 1. It reveals that 70% of calls experienced minimal waiting, while approximately 10% of customers had to endure waiting times exceeding 60 seconds during the overloaded periods.

Furthermore, due to the abandonment rate being much lower than the service rate (indicating that customer patience, measured as the time to abandon, is usually much higher than the average service time), the equilibrium of an overloaded system would be reached at a very high system state. As showed in our earlier example of an overloaded system, this would lead to unacceptable levels of abandonment. In reality, the system rarely reaches equilibrium. The real abandonment rate is a result of two phases during the overloaded period: first, the number of queued calls increases during the understaffed period and then decreases once the number of agents becomes sufficient to ensure a high service level. Therefore, the overloaded periods are typically transient, providing another argument against approximating them using stationary results of queuing models.

The inadequate number of agents during certain planning periods can occur due to intentional scheduling decisions, such as restrictions imposed by labor laws, or as a random result of external factors like weather conditions. These factors should be taken into account in the planning process. The model should be able to accurately estimate the service level and abandonment rate by considering both the planned overloaded periods (based on the forecasts) and the possibility of certain percentages or specific periods of time being overloaded.

Using the same "average" parameters in our approximate model for this distinct modes of the system would lead to a significant overestimation of the abandonment rate during periods of expected demand (QED periods) and would produce unsatisfactory results for both the SL and abandonment rate during overloaded periods.

To address this issue, a practical solution is to utilize two different model settings for the distinct modes of operations of the system, namely the QED mode and the overloaded mode e.g., when real-time adjustments to the number of available servers are no longer possible, the settings for the overloaded system should be employed to calculate the subsequent periods.

Another ultimate solution would involve the utilization of an inhomogeneous CTMC model that incorporates a precise depiction of the empirical abandonment distribution. One possible approach is to employ phase-type Coxian distributions, as proposed by Creemers et al. (2014). However, the computational complexity of resulting CTMC models poses a practical challenge, especially when repeated calculations are required during the optimization process, such as constructing schedules based on forecasts. Other proposals, as previously mentioned state-dependent Markovian birth-and-death queuing models, provide parameter fit methods for stationary systems only, limiting their applicability in this

context.

CONCLUSION AND FUTURE WORK

In this paper, we have presented a novel methodology for parameter fitting in a time inhomogeneous CTMC model of a call center, using real call center data. Our findings indicate that abandonment in real systems primarily occurs due to the inability of real-time adjustments during certain periods of the planned schedule. These periods can be a result of intentional schedule planning or temporary external events that lead to different call volumes than originally forecasted.

In particular, we have demonstrated that popular models, such as the Erlang-A model, which assume exponential distribution for abandonment, only provide accurate results during overloaded periods. In periods where the incoming call volume is as expected, models without abandonment yield more precise results.

In this particular case our methodology provides a practical solution for parameter fitting and suggests the use of different model settings for different modes of operation, enabling improved accuracy in forecasting and optimization processes.

Our approach has the capability to be extended for fitting other model parameters than the abandonment process. For example, it is commonly known that the exponential distribution is merely an approximation of the empirical service time distribution, which can be more accurately represented by lognormal distributions. Consequently, we can expect to observe similar discrepancies as in the case of mismatched abandonment distributions. Hence, our proposed method will probably provide a better fit, regarding the calculated performance measures, compared to simply using the average service rate obtained from real data.

The proposed methodology can be applied more broadly for modeling of other queuing systems where approximations with known deficiencies regarding their accuracy are used but offer e.g., computational advantages - for quickly determining their parameters. In particular, in our example, instead of using average values for the abandonment rate, we directly matched the results of the model, resulting in higher accuracy.

Additionally, the method enables to easily assess the range of situations where the chosen approximation would deliver correct results, such as the two distinct modes of operation in a call center - the QED and overloaded states.

By incorporating the empirical distribution of abandonment using phase type Coaxian distributions, future research could explore more sophisticated modeling techniques that capture the complexity of abandonment dynamics in call centers. Alternatively, the development of state-dependent Markovian birth-and-death queuing models that can handle non-stationary systems would also improve available modeling capabilities.

From the practical point of view, our methodology offers valuable insights and ready to use solutions for parameter fitting in call center models, paving the way for improved accuracy and effectiveness in call center operations and planning.

REFERENCES

Aksin Z.; Armony M.; and Mehrotra V., 2007. *The Modern Call Center: A Multi-Disciplinary Perspective on Operations Management Research. Production and Operations Management*, 16, no. 6, 665–688. doi:10.1111/j.1937-5956.2007.tb00288.x.

Baccelli F. and Hebuterne G., 1981. *On queues with impatient customers*. Ph.D. thesis, INRIA.

Brandt A. and Brandt M., 2002. *Asymptotic results and a Markovian approximation for the M(n)/M(n)/s+GI system. Queueing Systems*, 41, no. 1/2, 73–94. doi: 10.1023/a:1015781818360.

Burak M.R., 2018. *Efficiency Improvements to Uniformization for Markovian Birth-and-Death Models.* In *2018 23rd International Conference on Methods & Models in Automation & Robotics (MMAR)*. IEEE, 741–746. doi:10.1109/MMAR.2018.8486075.

Burak M.R. and Korytkowski P., 2020. *Inhomogeneous CTMC Birth-and-Death Models Solved by Uniformization with Steady-State Detection. ACM Trans Model Comput Simul*, 30, no. 3. ISSN 1049-3301. doi: 10.1145/3373758.

Cezik M.T. and L'Ecuyer P., 2008. *Staffing Multiskill Call Centers via Linear Programming and Simulation. Management Science*, 54, no. 2, 310–323. doi:10.1287/mnsc.1070.0824.

Creemers S.; Defraeye M.; and Nieuwenhuyse I.V., 2014. *G-RAND: A phase-type approximation for the nonstationary queue. Performance Evaluation*, 80, 102–123. doi:10.1016/j.peva.2014.07.025.

Defraeye M. and Nieuwenhuyse I.V., 2015. *Staffing and scheduling under nonstationary demand for service: A literature review. Omega.* doi:10.1016/j.omega.2015.04.002.

Deslauriers A.; L'Ecuyer P.; Pichitlamken J.; Ingolfsson A.; and Avramidis A.N., 2007. *Markov chain models of a telephone call center with call blending. Computers & Operations Research*, 34, no. 6, 1616–1645. doi:10.1016/j.cor.2005.06.019.

Erlang A.K., 1917. *Solutions of some problems in the theory of probabilities of significance in automatic telephone exchanges. Electroteknikeren, 13: 5–13, 1917. Danish [English transl in: PO Elec Eng J 10 (1917–18) 189–197.*

Feigin P., 2006. *Analysis of customer patience in a bank call center*. In *Working paper, The Technion*.

Green L.V.; Kolesar P.J.; and Soares J., 2001. *Improving the Sipp Approach for Staffing Service Systems That Have Cyclic Demands*. *Operations Research*, 49, no. 4, 549–564. doi:10.1287/opre.49.4.549.11228.

Ingolfsson A.; Campello F.; Wu X.; and Cabral E., 2010. *Combining integer programming and the randomization method to schedule employees*. *European Journal of Operational Research*, 202, no. 1, 153–163. doi:10.1016/j.ejor.2009.04.026.

Jouini O.; Akşin Z.; and Dallery Y., 2011. *Call centers with delay information: Models and insights*. *Manufacturing & Service Operations Management*, 13, no. 4, 534–548.

Jouini O.; Koole G.; and Roubos A., 2013. *Performance indicators for call centers with impatient customers*. *IIE Transactions*, 45, no. 3, 341–354. doi:10.1080/0740817x.2012.712241.

Kanavetas O. and Balcıoglu B., 2022. *The 'sensitive' Markovian queueing system and its application for a call center problem*. *Annals of Operations Research*, 317, no. 2, 651–664. doi:10.1007/s10479-018-2802-6.

Kim S.H. and Whitt W., 2013. *Statistical analysis with Little's law*. *Operations Research*, 61, no. 4, 1030–1045. doi:10.1287/opre.2013.1193.

Mandelbaum A. and Zeltyn S., 2013. *Data-stories about (im)patient customers in tele-queues*. *Queueing Systems*, 75, no. 2-4, 115–146. doi:10.1007/s11134-013-9354-x.

Palm R.C.A., 1957. *Research on telephone traffic carried by full availability groups. Tele (English translation of results first published in 1946 in Tele, vol7)*, 1, 107.

Petropoulos F. and et al., 2022. *Forecasting: theory and practice*. *International Journal of Forecasting*, 38, no. 3, 705–871. ISSN 0169-2070. doi:https://doi.org/10.1016/j.ijforecast.2021.11.001.

Whitt W., 1999. *Improving Service by Informing Customers About Anticipated Delays*. *Management Science*, 45, no. 2, 192–207. doi:10.1287/mnsc.45.2.192.

Whitt W., 2005. *Engineering Solution of a Basic Call-Center Model*. *Management Science*, 51, no. 2, 221–235. doi:10.1287/mnsc.1040.0302.

THE DATA USED IN THE BANKING CALL CENTER EXAMPLE

In this section we provide more details about the estimation of the parameters $\lambda(t)$, $s(t)$, $\mu(t)$, η and γ, based on true data from an existing call center.

The real data that we have used was extracted from a relational database made available by Mandelbaum and Zeltyn (2013). It contains calls collected in the period 2001 to 2003 at the telephone call center of a mid sized American bank (*dataset labeled U.S. Bank*). The call center had four sites: in New York, Pennsylvania, Rhode Island, and Massachusetts, integrated into a single virtual call center. The call center capacity was about 900-1200 scheduled agents on weekdays and 200-500 on weekends. The agents were assigned to different customer classes (skills) and the incoming calls from different customer types were assigned to particular agents using skill-based routing (SBR). Because our model represents a homogeneous (single skill) call center, we had to select a single customer class served possibly with a group of agents assigned to a single skill.

In particular, we used the data of the skill *Summit* (*service*=14) similarly as in Kim and Whitt (2013) and described in detail in their supplementary material on the call center data. A detailed description of the database, including the information contained in the various tables and an explanation of the different values of their fields can be found at:

https://see-center.iem.technion.ac.il/databases/USBank/Data/USBank.pdf

The working hours of the *Summit* line were between 6:00:00 and 22:59:59. Outside of this time no incoming calls were routed to the skill. However there were examples of the line working for shorter times - for example on 2001-06-19 all agents logged out already at 17:50 but the system continued to route customer calls to the line resulting in all calls abandoned. As we were particularly interested in observing of the abandonment behavior, we included in our model all time periods between 0:00:00 and 23:59:59 divided in small (e.g. 1440 of one-minute) periods, to be able to observe uncensored abandonment behavior due e.g., to misconfiguration of the system.

The incoming call data was extracted from the *calls* and *cust_subcalls* tables. From the *calls* table only the incoming calls served by *Summit* skill were selected (*entry_service = 14*). Additionally only the calls from the period between 2001-04-01 00:00 and 2001-07-25 23:59:59 were selected i.e., from the time, where the skill *Summit* was served continuously. Outside of this time period there were cases where the skill *Summit* was used (calls were routed into) but it was either served not continuously or the volumes of calls were much different then in the above period - probably the skill was "reused" temporarily for some other service.

In the mentioned period there were 427000 incoming calls:

month	no of calls
April	99268
May	122948
June	124909
July	73574

Every incoming call registered in the *calls* table got unique *call_id* number. A single call usually consisted

cust_subcall	record_id	outcome	wait_time	queue_time	service_time	hold_time	party_answered	agent_group	main_service
1	243222	20	4	0	0	0	0	0	0
1	243223	22	1	1	1545	459	23007	43	14
2	243227	22	0	0	103	96	23016	43	14
3	243228	1	0	0	1663	0	725	0	0

Table 1: Subcals resulting from call nr 485159427

of more then one subcalls corresponding to interactions with different parties. For example the call *485159427* consisted of 4 subcalls: first the customer interacted with VRU (the first subcall ending with the *outcome = 20*) then he was put into the queue and after the waiting time of 1 s. the call was answered by an agent belonging to the *Summit* skill, by whom the call was handled for 1545 s. During this time the customer was put on hold for 459 s and transferred to another agent from the skill *Summit* (*outcome = 22*). The resulting subcall (*record_id = 243227*) was again transferred to another agent, this time to one not belonging to the *Summit* skill, resulting in a new subcall (*record_id = 243228*) which was finished by the customer (*outcome = 1*). For the purpose of our model we treated calls forwarded to other skills as calls leaving the system and consequently changing (decreasing) the state of the system. Consequently transfers inside the *Summit* skill do not change the system state but increase both service rate and the number of incoming calls. In case of re-entry (transfer from other skills) we considered them in relation to the system state (increased number of calls in the system and number of busy agents), the service rate and number of incoming calls. Because the transferred call does not experience waiting i.e., it is put on hold and transferred when an agent from target skill becomes available and therefore is already receiving service, we did not account for such calls regarding their patience/abandonment behavior.

For the estimation of the number of incoming calls in a analyzed period ($\lambda(t)$) we included all subcalls entering the system, including subcalls transferred from other agents. In order to simplify other database queries all such calls were selected via database view *summit_subcalls*. The service rate $\mu(t)$ in a period included all subcalls finished by an agent either due to customer hanging up (*outcome = 1*), agent hanging up (*outcome = 2*) or agent transferring customer (*outcome = 22*) to another party (another agent or VRU). The true system state used to compare with the outcome of the model includes all calls handled by agents belonging to the skill *Summit* at the end of a period (*time_period LEFT JOIN summit_subcalls ON summit_subcalls.segment_start<=time_period.per_end AND summit_subcalls.segment_end>time_period.per_end*).

The table *time_period* contains timestamps of beginning and end of all analyzed time periods in the format used for the timestamps in the tables *calls*, *cust_calls* and *agent_events* i.e., number of seconds since 1970-01-01 00:00. The lengths of the periods can differ. For the purpose of the connected papers they were all set either to 60s or 5min.

The distribution of customer patience (time to abandon) and the preliminary values of balking rate γ and abandonment rate η of the model, were estimated only for subcalls queued after exiting VRU (without transfers as already explained) for uncensored periods only i.e., in order to estimate the distribution of times to abandon smaller than X s we considered only the time periods where minimal waiting time of answered calls (censoring limit) was greater than X s. We define an abandoned call as a subcall finished by the customer before receiving the service (*outcome IN (11,12,13)*) or during the transfer to the agent (*outcome=1 AND service_time < 2*) similar to the definition of abandonment used by Feigin (2006). For the subset of abandoned calls defined as balking cals we did not rely on the outcome assigned in the (sub)call records (*outcome=11*). Instead an arbitrary waiting time limit was used.

The estimation of the number of available agents *s(t)* used data from the table *agent_events*. We considered only agents either logged in to the skill *Summit* directly (after an agent event with *event_id=20*) or logged in from other skill (*event_id=21*). In both cases the agents were receiving calls from the queue and via transfers from other agents. If an agent served only transferred *Summit* calls (*service = 14*) e.g., an escalation to a supervisor or transfer to another skill, then we did not consider such agents as part of the server pool in our model and treated the transferred subcalls as leaving the system, as already mentioned above. Beside of login into and logout of the skill, an agent can also change status to break. Following the analyses of the agent events and call records, the authors presume the break with the *event_id=62* being a 'private' break and the break with the *event_id=61* a break related to the work e.g., some offline activities like answering emails, data completion etc. Because the duration of agent records connected to agent events related to handling calls can exceed the time spent on the interaction with the customer (*talk_time + hold_time*) i.e., includes additional *wrapup_time*, we assume that all additional activities that are related to a call are handled completely during agents events related to calls. Consequently we exclude agents being on breaks (both *61 and 62*) from available agents. Additionally, an agent is considered as available for a given period only when being available for $\geq 50\%$ of the period time.