Chat Based Contact Center Modeling
System Modeling, Parameter Estimation and Missing Data Sampling

Per Enqvist[1] and Göran Svensson[1,2]
[1]*Division of Optimization and Systems Theory, Kungliga Tekniska Högskolan, Lindstedtsv 25, Stockholm, Sweden*
[2]*Teleopti WFM, Teleopti AB, Stockholm, Sweden*

Keywords: Queueing, Chat, Chat-based Queueing Systems, Parameter Estimation, Gibbs Sampling.

Abstract: A Markovian system model for a contact center chat function is considered and partially validated. A hypothesis test on real chat data shows that it is reasonable to model the arrival process as a Poisson process. The arrival rate can be estimated using Maximum likelihood. The service process is more involved and the estimation of the service rate depends on the number of simultaneous chats handled by an agent. The estimation is made more difficult by the low level of detail in the given data-sets. A missing data approach with Gibbs sampling is used to obtain estimates for the service rates. Finally, we try to capture the generalized behaviour of the service-process and propose to use generalized functions to describe it when little information is available about the system at hand.

1 INTRODUCTION

Contact centers usually offer several types of media to enable customer communication. Chat functionality is one such type of media that in recent years has grown in popularity. This stresses the importance of good modeling and parameter estimations for chat based systems.

In this paper a Markovian chat system model is considered (Enqvist and Svensson, 2017). The chat system is viewed as a queue-based state-space model, akin to traditional queueing systems of inbound telephone call centers, described in detail elsewhere such as in (Koole, 2013; Gans et al., 2003; Aksin et al., 2007). However, chat systems behave slightly different than traditional telephone queues in that an agent can work with several customers simultaneously. We make the assumption that the number of customers an agent is serving effects the service rate with which the service is provided. A queueing based state-space model should capture both how many customers each agent is currently serving and how many customers there are in the system in total, as well as the varying service rates.

The main goals of this paper are to argue that such a queueing-based state-space model is reasonable, to support such a model by use of real world data and to propose methods for estimating the rate parameters for a chat system from real but incomplete data-sets.

The quality and level of detail of the underlying data can severely limit the choice of methods and the uncertainty of estimates. When there exist limitations on available data it may be necessary to rely more on prior information and thus we propose that general parameterized functions are used to lower the variance of the estimates by including empirical information about chat systems general behaviour.

Due to the strong dependence on data we propose a data classification structure that pertains to chat systems. The classification of underlying data would indicate which technique is appropriate. The accuracy and choice of any, realistic, model will be data dependent. For a statistical analysis of a call center see (Brown et al., 2005).

It is natural to divide the problem into two parts, one part for the arrival rate process and one for the service rate process. The former lends itself to standard methods of estimation when the level of detail of the underlying data is good enough. While the latter often leads to complications due to data on aggregated or mean value form, *i.e.*, low level of detail.

We show that it is reasonable to describe the arrival process in terms of a homogeneous Poisson process on 15-minute or 30-minute intervals. We support this position by χ^2-square hypothesis testing at five percent significance level on two data-sets. For more on nonhomogeneous arrival processes we refer to (Green and Kolesar, 1991) and (Whitt, 2007).

Enqvist P. and Svensson G.
Chat Based Contact Center Modeling - System Modeling, Parameter Estimation and Missing Data Sampling.
DOI: 10.5220/0006251604640469
In *Proceedings of the 6th International Conference on Operations Research and Enterprise Systems (ICORES 2017)*, pages 464-469
ISBN: 978-989-758-218-9

When estimating at the service rate process the situation gets more complicated. One possible cause of complications occur if the start and finish times of chat dialogues are not recorded. In our data-sets only the number of initiated chats per agent and interval is available. Since an agent can serve several customers in parallel we make the assumption that the service per customer is a non-increasing function in the number of currently served customers. In (Bekker et al., 2004) the authors explore varying service levels and in (Bekker et al., 2011) adapting service rates are investigated. In computer systems processor sharing is a common phenomena, see (Cohen, 1979). Our model is inspired by both the previous situations, where an agent has capacity to perform simultaneous tasks but at varying rates. A further complication is due to data often only being available on an aggregated level. Thus it is not possible to discern the actual (pointwise) workload distribution for the interval. We suggest a missing data approach, via the expectation maximation algorithm and Gibbs sampling, to handle this problem.

There might also exist general information about the system, such as how likely it is that there are customers waiting in the queue and the arrival rate from a previous estimation. One might also include data from other chat systems and assume that there are similarities. Hence we propose to model service rate per customer as a continuous non-increasing function, depending on the state of the chat system and the specific agent. Such a function can provide answers about the maximum allowed chats in parallel to fullfill some quality of service goal, like maximizing throughput through the system or to support staffing decisions.

In Section 2, data is discussed and the data-sets are presented. In Section 3, the proposed queueing-based state-space model is introduced and parametrized. The parameters to be estimated are also stated. In Section 4, the estimation models for the arrival process is explained and the hypothesis testing is show for specific data-sets. Also the missing data approach for estimation of the service rates is presented.

2 DATA CHARACERISTICS

What can be achieved in terms of reliable estimates, in a contact center environment, is highly dependent on the amount and quality of the available data. Therefore, it makes sense to categorize data in terms of quality. We identify three major aspects that determine the overall quality and three subsets that are important for estimations in queueing systems, namely: 1. Number of data records, 2. The level of detail, 3. Relevant data-sets. The data-sets can be split into general system, agent specific and customer specific data.

The number of data records is an important factor in determining the level of accuracy of estimates. The level of detail determines how easily one can perform estimations. Furthermore, in the context of queueing systems, it is meaningful to differentiate between three types of data-subsets. The first set concerns data on a system level, such as offered load per interval. The second subset pertains to agent specific data, data like agent-id and number of initiated chat dialogues per interval. The third subset of data records contain information on individual customers, such as customer-id, arrival time to the system and waiting time in queue.

In cases where there are few data records, low detail level or when not all three subcategories are available leads to uncertainty in the estimations. This type of uncertainty has to be managed, which motivates why we need methods to provide reliable estimates in the face of poor data quality.

The given data-sets, on which this paper is based, come in two subsets, where the first subset contain general queue data and the second contain agent specific data. Thus customer specific data is missing in all cases. The data deemed useful in the context of this paper is presented, while other data posts not deemed to influence the procedings is supressed.

After discussing the matter with responsible data base administrators it is found that the data is not completely machine generated and thus may contain errors due to human factors. This type of problem requires serious attention but for the purposes of this text it is ignored apart from some pre-processing with respect to outliers and records with low information content.

2.1 General Queue Related Data

In the first type of data subset the important data posts are the ones representing date, intraday intervals and offered load. The data is given per date and per interval, thus we introduce $d \in \mathcal{D} = \{1,\dots,D\}$ indexing the days, $i \in \mathfrak{I} = \{1,\dots,I\}$ indexing the intraday intervals and $w_d \in \{1,\dots,W\}$ index the day of the week, where $W = 7$. Let $N_{d,i} \in \mathbb{N}$ represent the number of arrivals on day d in interval i, *i.e.*, offered load. The notation was inspired by (Gans et al., 2009).

2.2 Agent Specific Data

In the second type of data subset the important data posts are the ones representing date, intraday interval, agent identity, number of initialized chats and aggregated time spent with open chat dialogues. The time spent chatting is the sum of all chat dialogues, thus the total time can be greater than the length of the corresponding interval.

2.3 Customer Specific Data

Many contact centers neglect to record customer specific data, such as arrival time, time waiting, time in service, service provided by agent-id, time of abandonment, etc. When such level of detail exists it is straight forward to estimate service rates and related parameters. The customer specific data-sets are missing for the chat systems investigated in this paper.

2.4 Given Data-Sets

Information of given data-sets. The size is measured as the raw data text file size.

Table 1: Given data-sets.

Data-set	Syst. data	Ag. data	Size
TA: queue	yes	no	5.3 MB
TA: agent	no	yes	49.7 MB
TC: queue1	yes	no	3.5 MB
TC: queue2	yes	no	2.3 MB
TC: agent	no	yes	8.4 MB

Where TA is a large travel agent company and TC is a large telecom company. Syst. data is short for System data and Ag. data is short for Agent data.

3 STATE-SPACE QUEUEING MODEL DESCRIPTION

The queue-based system considered here is approximated by a Markovian state-space model that is described in detail in (Enqvist and Svensson, 2017). The states represent the total number of customers in the system, the number of agents working during that interval and also contains information about the number of customers that each agent is currently serving, possibly up to some maximum number. We refer to (Asmussen, 2003) for queueing theory in general.

State transition rates are determined by new arrivals and completed service sessions. New arrivals are either placed in a common queue or start recieving service from an available agent according to a routing rule.

Completed services depend on the number of agents, the number of customers in parallel that each agent is serving and the corresponding service rates according to a generalized processor sharing model, as in (Cohen, 1979), for each agent.

An example, of a chat queue with one agent and a maximum of three jobs in parallel, is shown in Figure 1. The lower chart shows the jobs as they are seen by the agent. When a new job starts recieving service the number of customers waiting in line, if any, drops by one, as seen in the top figure. The example can be expanded to include more agents which, together with exponential service times and Poisson arrivals, quite naturally gives rise to a Markovian state-space model.

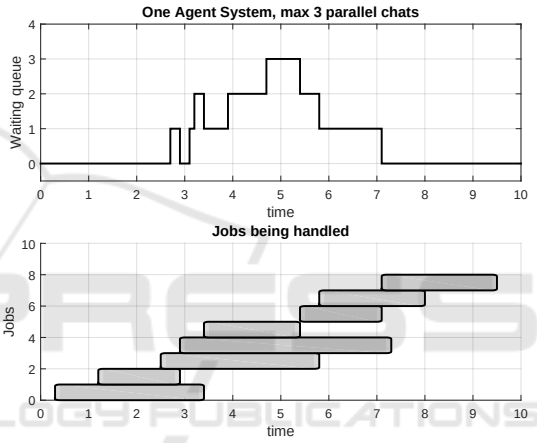

Figure 1: An example of a one agent chat queue. In the upper graph the number of customers currently waiting in queue is depicted. In the lower chart the different jobs are shown, from the time they start recieving service until they leave the system. The maximum number of customers that can be served in parallel, in this example, is three.

3.1 The Arrival Process

New customers enter the system according to an arrival process that is assumed to be Poisson with parameter $\lambda_{w_d,i}$, where w_d correspond to the day of the week and i to the interval. We assume that arrivals are independent and identically distributed. That this is a reasonable assumption is demonstrated in Section 4.1, as part of the model validation process.

Newly arrived customers either have to wait in line or are assigned an agent according to the routing rule. In this paper they are routed to the agent with the least number of customers in service. In the case that there are more than one agent serving the same, least number, of customers then random selection is used.

3.2 The Service Process

The service rates of the agents can be studied per agent, by groups of agents or by assuming all agents perform at the same rate of service. The two former situations require more data and can cause the underlying state-space to grow extremly large, hence in this paper we choose to treat all agents as equivalent. It might also be considered that agents provide the same kind of service during different intraday intervals or it may be assumed that the service rates vary depending on the interval. The latter case requires more data and may be impractical due to the need to solve many versions of the system.

The state of an agent will be taken as the number of currently served customers in parallel, possibly up to some maximum number.

Furthermore, we assume that the service times of the agents are exponentially distributed with rate parameters μ_j, where $j \in \{0,1,\ldots,m\}$ represents the number of currently served customers. These service rates are dependent on the state of the agent and represent the fact that an agent serving several customers simultaneously cannot devote the same type of attention to all of them as to a single customer.

The service times of the agents are assumed to be independent of each other and identically distributed, only depending on the number of customers in parallel. Here we assume that the different customers served by the same agent at the same time are independent.

In Section 4.2, we consider the estimation of the service rates of the agents.

4 PARAMETER ESTIMATION AND MODEL VALIDATION

This section will be divided into two major parts, the first account for the arrival process and the validity of the Poisson assumption and the second part considers the service process and the estimation of the service rates for the agents, subject to varying numbers of chats in parallel.

4.1 The Arrival Process

Data is given on the form of date, 15- or 30-minute interval and offered load, *i.e.*, number of arrivals $N_{d,i}$ per day and interval. We collect all data-points pertaining to each day of the week and intraday interval in a vector, denoted by $\tilde{a}_{w_d,i}$ of length $L_{w_d,i}$.

All the $\{\tilde{a}_{w_d,i}\}$:s are then pre-processed by removing outliers and entire vectors that contain too little information. Let $\{\bar{a}_{w_d,i}\}$ denote the resulting processed set of vectors, now containing only trusted data. Assuming that the arrivals are independent and identically distributed and specifying the likelihood function to be the joint function for all observations in the specified vector $\bar{a}_{w_d,i}$

$$\mathcal{L}(\lambda_{w_d,i};\bar{a}_{w_d,i}) = f(a_1^{w_d,i},\ldots,a_{L_{w_d,i}}^{w_d,i}|\lambda_{w_d,i}). \quad (1)$$

$L_{w_d,i}$ varies from vector to vector depending on the given data-set. Then a maximum likelihood estimation (MLE) of the Poisson parameter can be performed to obtain the corresponding arrival rate per interval and day of the week. The estimation is unbiased and given by the sample mean (Haight, 1967, Ch. 5)

$$\hat{\lambda}_{w_d,i} = \frac{1}{L_{w_d,i}} \sum_{l=1}^{L_{w_d,i}} a_l^{w_d,i}. \quad (2)$$

An example of estimated arrival rates for a chat system of a travel agency are shown in Figure 2. The actual arrival rates have been modified by request from the company, but the general behaviour is captured. It is shown for 30-minute intervals.

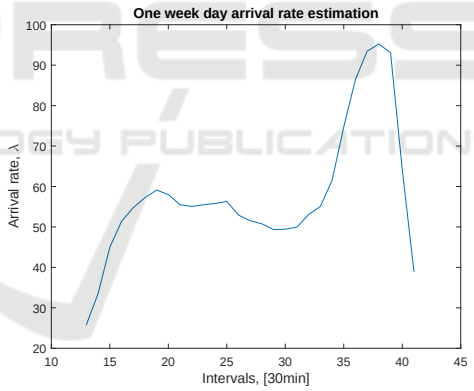

Figure 2: MLE of the arrival rates for one day of the week with half hour intervals. (Rates are modified by request from the company).

A common assumption for many queueing systems is that arrivals can be modeled by a homogeneous Poisson process. To show that this is plausible the data is tested via a Pearson χ^2-test, with test statistic given by

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \quad (3)$$

where n represents the maximum observation value category, O_i the number of observations of type i and E_i the expected number of observations of type i. The

null hypothesis, H_0, is defined to be that data is Poisson distributed with parameter $\hat{\lambda}_{w_d,i}$, and the test is performed at 5% significance level.

Examining the data-sets it is found that the assumed Poisson arrival rate is not rejected for the majority of the intervals and day of the week, results given in Table 2.

Table 2: Example showing the number of non-rejected and rejected H_0-hypothesises for two different pre-processings of the same 15-minute interval data-sets.

Data	Pre-process	Not rejected	Rejected
TA	Low	277	110
TA	High	260	52
TC2	Low	232	210
TC2	High	232	46

The intervals for which the Poisson assumption is rejected are mostly found in the beginning and at the end of the day. When the data is aggregated into half hour intervals the frequency of rejections decrease. In Table 2 it can be seen that the result is dependent on the pre-processing of the data, thus non-automated data processing was needed to obtain the results. Information to this end was supplied by the data-base administrators.

Considering the results we find it reasonable to model the arrivals as a Poisson process for the intents and purposes of this work, for a similar but detailed paper see (Brown et al., 2005)

4.2 The Service Process

The queue and service are modelled as a Markov process, where the service rates of each agent depends on the number of current clients. To decide if this model assumption is reasonable we would like to perform a hypothesis test. However, such a test could be performed in the full information case, where client data is available, but for the given data sets such a test is not easily performed.

To illustrate the difficulties arising from data on an aggregated format we introduce an example case as can be seen in Figure 3, were the cumulative arrivals and the cumulative number of answered chats is shown. Note that only the interval in which arrivals and answered chats occur is known.

Assuming that the model setup is suitable, we concentrate on the problem to estimate the various service intensities of the agents.

For the full data case the intensities can be estimated from information about the state transitions of the system, by using a MLE method. However, the given data-sets lack this level of detail and a direct MLE is not feasible.

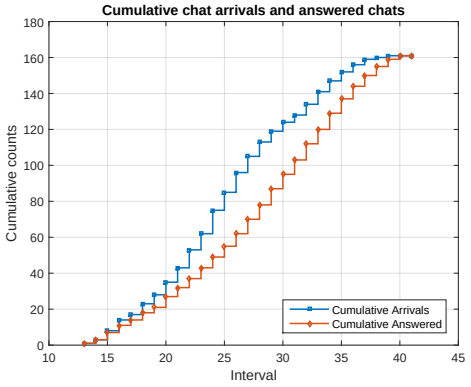

Figure 3: Example showing the cumulative number of arrivals and answered chats for a chat queue where data is aggregated per interval.

One approach is to try to estimate the missing data first and then apply the MLE approach. Consider one agent over a sequence of time windows $i = 1, \cdots, n$ in $\mathfrak{I}$. Let x_i denote the information about block i needed to determine the MLE, and let y_i denote the information about block i that is observable, *i.e.*, the information provided by the considered data set. Assume that θ contains all the unknown intensities μ_j. It would be very difficult, or even impossible, to determine a closed form expression for the probability of observing the provided data given the intensities θ, *i.e.*, for $\Pr_\theta(y_{1:n})$. Here, $y_{1:n}$ denotes the values of y_1 to y_n. Therefore, we propose to apply the expectation maximization (EM) algorithm (Moon, 1996) and (Dempster et al., 1977) to determine the estimate of θ. Starting from an initial guess θ_0 we want to determine

$$Q(\theta;\theta_0) = \mathbb{E}_{\theta_0}\left[\log \Pr_\theta(x_{1:n}, y_{1:n}) | y_{1:n}\right], \qquad (4)$$

for the expectation step of the EM algorithm. To take the expectation under θ_0 it is necessary to have $\Pr_{\theta_0}(x_{1:n}, y_{1:n})$. It can be determined using single site Gibbs sampling (Gelfand and Smith, 1990) where we should make use of the given observable data. Sequentially determine $\Pr_{\theta_0}(x_i | x_{1:i-1}, x_{i+1:n}, y_{1:n})$ for each $i = 1, \cdots, n$. If $y_{1:n}$ contains information about the number of arrivals and working time in the blocks it should be used to determine the probabilities. Thus we need the x_i:s to contain the missing information needed to determine the intensities. Let $x_i = \{d_i, B_1, \ldots, B_{d_i}, T_1, \ldots, T_{\ell_i}\}$, where d_i represents the number of finished chats in interval i, B_j are the inter-departure times, T_j the inter-arrival times and ℓ_i the number of new arrivals in interval i, which is observable. The maximum number of customers in block i is bounded by the number of customers at the beginning of the interval, N_i, and the number of arrivals during the interval ℓ_i. Let $z_i = \{x_{1:i-1}, x_{i+1:n}, y_{1:n}\}$, then the

conditional probabilities

$$\Pr(x_i|z_i) = \sum_{j=1}^{N_i+\ell_i} \Pr(x_1|z_i, d_i = j) \Pr(d_i = j|z_i), \quad (5)$$

can be determined. These correspond to all possible combinations of jumps in an interval.

The final piece is to use that we can observe the total chat time in interval i. This impose a linear equality constraint on the variables in x_i.

In the maximization step of the EM algorithm an updated estimate θ_1 is determined from

$$\theta_1 = \underset{\theta}{\operatorname{argmax}}\, Q(\theta; \theta_0). \quad (6)$$

This process is repeated until sufficient convergence has been achieved.

4.3 Proposed Service Rate Function

In order to reduce the number of parameters to estimate we propose that a parametric function representation of the service rate is used. This function class can be chosen to represent physical properties of the rate parameters. We propose the following function class.

$$f(\tilde{n}) = \begin{cases} 0, & \tilde{n} < 1 \\ \tilde{n}a\left(1 - \frac{1}{1+be^{c(d-\tilde{n})}}\right), & \tilde{n} \geq 1 \end{cases} \quad (7)$$

where $\tilde{n} \in \mathbb{R}$ is the continuous version of the number of customers per agent, and a, b, c and d are nonnegative model parameters. The function captures the desired shape but fitting the parameters from data is not trivial. With this representation we ensure that the service rate per customer is nonincreasing.

5 CONCLUSIONS

We have shown that it is reasonable to model the arrival process as a Poisson process via hypothesis testing. An approach for estimating the service rate parameters from observed data has been proposed.

Implementation and further validation of the model and estimation procedure is currently in process. Other further work would be to use an alternative Bayesian approach.

ACKNOWLEDGEMENTS

This work has been made possible by Teleopti AB, both financially and by providing data. Thanks go to T. Pavlenko, J. Olsson and F. Rios for valuable input.

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, Vol. 16, Issue 6.

Asmussen, S. (2003). *Applied Probability and Queues*. Springer-Verlag, New York, 2nd edition.

Bekker, R., Borst, S., Boxma, O., and *et al.* (2004). Queues with workload-dependent arrival and service rates.

Bekker, R., Koole, G., Nielsen, B., and *et al.* (2011). Queues with waiting time dependent service.

Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S., and Zhao, L. (2005). Statistical analysis of a telephone call center. *Journal of the American Statistical Association*, 100(469):36–50.

Cohen, J. W. (1979). The multiple phase service network with generalized processor sharing. *Acta Informatica*, 12(3):245–284.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38.

Enqvist, P. and Svensson, G. (2017). Chat based queueing systems with varying service rates and simultaneous jobs. To be submitted.

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

Gans, N., Shen, H., Zhou, Y.-P., Korolev, K., McCord, A., and Ristock, H. (2009). Parametric stochastic programming models for call-center workforce scheduling. *Technical report*.

Gelfand, A. E. and Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, 85(410):398–409.

Green, L. and Kolesar, P. (1991). The pointwise stationary approximation for queues with nonstationary arrivals. Management Science, 1991, Vol.37(1).

Haight, F. (1967). *Handbook of the Poisson distribution*. Publications in operations research. Wiley.

Koole, G. (2013). *Call Center Optimization*. MG Books, Amsterdam, 1st edition.

Moon, T. K. (1996). The expectation-maximization algorithm. *IEEE Signal Proc. Magazine*, 13(6):47–60.

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