# A New Delay History Predictor for Multi-skill Call Center

Mamadou Thiongane(✉), Mohamed M. Ould Deye, Modou Gueye, and Mbaye Séne

Department of Mathematics and Computer Science, University Cheikh Anta Diop, Dakar, Senegal
{mamadou.thiongane,mohamed.oulddeye,modou2.gueye,mbaye.sene}@ucad.edu.sn

**Abstract.** We are interesting in methods for predicting the time that a customer must wait at this arrival in multi-skilled call centers. We propose a free-parameters delay history predictor that can be used in multi-skilled contexts. It only uses the waiting times of previous customers in the same class who found the same length of queue when they joined the system, and computes a weighted average of this past delays. In our numerical experiments with simulation models and real multi-skill call center, this new predictor is very competitive with existing delay history predictors. It gives often better results than other delay history predictors, and it is also very easy to implement in practice. This delay predictors can also be used in other service systems such as medical clinic, bank or emergency service.

**Keywords:** Delay History Predictor · Real Data · Call Center

## 1 Introduction

In service systems like emergency services, banks queueing services, call centers, and so on, announcing the estimated waiting time to new customers at their arrival can greatly increase customers satisfaction and experience at service providers. It can also help to improve the global system performance. By way of example, many hospitals in the USA and Canada periodically calculate the average waiting time for patients in their emergency rooms, and the results are published online or displayed on digital dashboard. This information can help to reduce emergency rooms saturation by encouraging new patients to go to a less crowded hospital. In call center, the queue is generally invisible to a calling customer, contrary to a physical queue in bank, supermarket or department store. The customer can only rely on the information provided by the call system. Providing delay announcements to customers, in addition to increasing customer's satisfaction, can significantly reduce the number of abandonments, and increases the overall system service rate. A growing number of call centers now provide delay announcement to their customers when they arrive in queue. Once the customer hears the estimated delay, she or he can choose to leave the


K. Cheikh M. F. Kebe et al. (Eds.): InterSol 2024, LNICST 610, pp. 139–150, 2025.
https://doi.org/10.1007/978-3-031-86493-3_12

queue, wait in the queue until receiving service or request to be recalled later if this option is available [2].

The related work on delay estimation can be classified in two categories. Some studies, like ours, focus solely on delay prediction without announcement to customers, whereas other studies integrate the impacts of the announcement of delays on the attitudes of the customers into their estimation models.

Most of the studies on delay prediction has been conducted for single queue systems. These works can be divided into two categories: *"Queue-Length"* (QL) delay estimators and *"Delay-History"* (DH) estimators. In QL delay estimators, the length of the queue and the parameters of the system as the number of used servers, and their service rate are used to predict the waiting delay, while the DH estimators use just the past customers delay time to predict the waiting time of a new customer in the system. The *"Last-to-Enter-Service"* (LES) predictor, which predict the waiting time of a new arrival customer by the wait time of the last customer who began his service, is the most popular DH predictor. DH predictors are much less accurate than QL predictors, however the latter are not applicable in a multi-skilled call centers. In modern call centers, which are multi-skill system, customers are classified by call type, and agents are grouped according to their skills. An agent can only serve a customer if she/he has the skills required for that call type. In Gan et al. [4] a well-detailed description of all the operational aspects of modern call centers has been done. In this work, the words "customer" and "call" are used interchangeably, as well as "server" and "agent".

Very few delay estimators have been done for multi-queue and multi-server systems, such as modern call centers. Senderovich et al. [11] have proposed predictors for a multi-skill system with only one agent group and many call type. Thiongane et al. [13,15,16] study more general method of delay estimation which can be applied in multi-skill systems. This work uses machine learning algorithms (Artificial Neural Networks (ANNs), Regression Smoothing Splines (RS)) and data collected on the system to learn a prediction function. The machine learning methods give good performance in multi-skill context but one drawbacks of them is they need many data and computational time to train model. The machine learning methods are also not easy to apply in practice, and one would have to rely on the simpler DH predictors.

In this paper, we propose a new DH delay predictor which predicts the customer waiting delay by an exponential smoothing weight average of the wait time of the past customers that have found the same queue length in the system at their arrival's. We call this predictor the *Weigth Exponential Smoothing Average Conditional LES* (WAvgC-LES). This predictor, like most DH predictors, are attractive in practice, because it requires no parameter estimation, and no optimization. We are studying these delay predictors in call centers context, but they could also be used in many other kinds of service systems.

In this work, we do not look at the influence of delay notification on the customer's waiting time. In practice, we often find the *LES* or *average LES* (Avg-LES). Thiongane et al. [14] propose AvgC-LES which is more accurate

than LES and Avg-LES but the latter need to store too many data to give good prediction unlike WAvgC-LES which stores only little data. In our numerical experiments with real and simulated data, we observe that WAvgC-LES is often more accurate than AvgC-LES.

The rest of this document is organised as follows. In Sect. 2, we present work that are done on delay estimation for service systems. In Sect. 3, we describe the general structure of modern multiskill call center models and present three examples of model which are used to evaluate the efficiency of predictors. In Sect. 4, we introduce the new delay predictor and also present other delay predictors that will be used to compare their performance with that of the new one. This comparison is made in Sect. 5. In Sect. 6, we conclude this work.

## 2 Review of the Literature

Most of work for delay estimation method has been done for single queue system for which customers are served in FCFS order. In that system a new arrival customer does not affect the waiting of customers that are already in queue. Assume a new customer who enters in a queue in which there are $K$ customers already waiting. Let $W$ denote the random variable representing the customer waiting time. A naturel and good predictor of $W$ is the QL predictor, which predicts its expectation conditional on $K$. For the GI/M/$c$ queue in which we have $c$ servers, arrivals follow a general distribution, service times follow an exponential distribution with mean $\mu^{-1}$, and customers have infinite patience (no abandonment) then the conditional expectation of $W$ is given by $\mathbb{E}[W \mid K] = (K+1)/(c\mu)$ [17]. In a GI/M/$s$+M queue, where customers have exponential patience time with rate $\nu$. The virtual expected delay, conditional on $K$, is predicted by $\mathbb{E}[W \mid K] = \sum_{k=0}^{K} 1/(c\mu + k\nu)$.

It is often very difficult to develop QL predictors for multi-skilled systems. For these system, we have multiple queues, agents have limited skills (i.e. each agent has a subset of customer types that he can serve), and agents often assign different priorities to different customer types. For the special case where each server can serve all customer types with the same order of priority, Senderovich [12] propose QL formulas for the special case that give upper and lower bounds on the expected delay time. Ibrahim et al. [7,9] propose DH predictors for single-queue systems. However, it should be noted that these predictors can be used in multi-skilled systems. These predictors use the waiting times of the past customers to predict the waiting time of a new customer. They include *"last-to-enter-service"* (LES) customer, *"head-of-line"* (HOL) customer, *"last-to-complete-service"* (LCS) customer, or the most *"recent-arrival-to-complete-service"* (RCS). The authors show that LES and HOL give more accurate prediction than RCS and LCS. Thiongane et al. [14] propose two DH predictors. The first is called E-LES. It estimates the waiting time of a new customer by extrapolating the waiting history of customers in the queue, and returns a weighted average of the extrapolated waits. The second, called AvgC-LES, predicts the waiting time of a new customer by averaging the waiting times of customers of

the same type (class) already served and having observed the same queue length on arrival.

Another class of predictors (usable in multi-skill context) which use machine learning algorithms (e.g., decision trees, splines regressions, and artificial neural networks) are proposed in recent years. The reader can see for example [1,10, 12,13,15,16] for more information on their implementation. These algorithms use data collected on system to learn the delay predictors. They give predictions that are better than those of DH. Their disadvantage is that they are complex to implement in practice. A large amount of data is required, and learning step times can often be too long. This is why simpler methods continue to attract interest.

## 3 The Model of Call Center

Here, we consider multi-skilled call center models. In these call center customers are classified according to the type of service they require. Agents are also divided into groups and the agents in the same group have the same skills. Call center opening hours are divided into periods of equal length. Customers arrival rate vary throughout the day. The distribution of the arrivals, the service time and the patience times can be very general. In the numericals examples with simulation models, we assume that arrivals of call type $k$ at period $p$ follow a Poisson process with rate $\lambda_{k,p}$ which is constant in period $p \in P$, service time are exponentials with rate $\mu_k$ and patience times also exponentials with rate $\nu_k$. A customer leaves the queue when his waiting time exceeds his patience time. We define by $s_g = (s_{g,1}, \ldots, s_{g,P})$ the staffing vector of agents in group $g$, where $s_{g,p}$ is the number of agents of group $g$ at period $p$. We use one queue per service type and customers in the same queue are served with *"first-come, first-served"* (FCFS) rule. If a call of type $k$ arrives and there are no agents available with the skills to serve it, it will be placed in the $k$ queue. Calls will be assigned according to the routing policy used in the examples.

Figure 1 shows three of the canonical models of multi-skilled call centers [5]. The first is the "V-model" with two types of calls and one agents group that handle the both types of calls. The second is the "N-model" with two type of call and two agent groups, where the group 1 have the skill sets to serve call type 1 et the group 2 can serve both call type. The third is the "W-model" with three calls types and two agents groups. The group 1 have the skill to serve call type 1 and 2, and the group 2 have the skill to serve the call type 2 and 3.

## 4 The Predictors of Delay

We start this section by presenting the DH predictor used in this study, and we finish by presenting our new DH predictor. For comparison, in our numerical experiment, we use a machine learning delay predictor (ANN) that is the better in multi-skill settings. It should be noted that, even though the ANN perform much better, these predictors have other drawbacks, as mentioned above. The

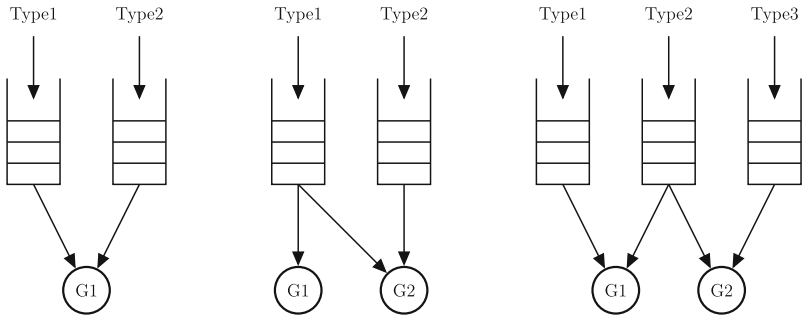


**Fig. 1.** The V, N, and W multi-skill models.

DH predictors use only the past customers delays of the same type to predict the waiting time of a new arrival customer in a queue. So, for each method considered, there is a different predictor for each call type class $k$, even if we don't always index it explicitly by $j$.

### 4.1 Last-to-Enter-Service (LES)

The predicted waiting time for a new customer is the wait time experienced by the last customer of the same type who was begin his service [8]. It is the most used DH predictor in practice.

### 4.2 Average-LES (Avg-LES)

This predictor is a version of LES that returns the average of several LES prediction. It predicts the waiting time of a new customer by averaging the waiting times of the $N$(A constant integer number) last customers of the same type who entered service. As LES, It is often used in practice [3]. Like LES, it is also widely used in practice.

### 4.3 Proportional-Queue-LES (P-LES)

The P-LES predictor is a predictor that takes into account the variation in queue length [6]. It uses the queue length $Q$ at the arrival of the new customer, the queue length at the LES customer's arrival $QL_{\text{LES}}$ and its waiting time $x$, and predict the new customer's waiting time $D$ by the ratio $Q/Q_{\text{LES}}$. The predictor actually used is :

$$D = x\frac{Q+1}{Q_{\text{LES}}+1}.$$

It solves the case where the last customer to enter service is to find an empty queue ($Q_{\text{LES}} = 0$).

### 4.4 Extrapolated-LES (E-LES)

To predict customers waiting time, the E-LES predictor uses information on the waiting delays of customers actually in the queue. The final waiting times of these customers (which are unknown) are obtained by extrapolating the times that they have already elapsed. E-LES returns the weighted average of the extrapolated waiting times as the prediction of the new customer's waiting time [14].

### 4.5 Average-LES-Conditional-on-Queue-Length (AvgC-LES)

This predictor, introduced by Thiongane et al. [14], predicts the waiting time of a new customer by averaging the waiting times of customers already served who have observed the same queue length on arrival as the new customer. The authors have shown in some simulated systems and in some real call centers data that this predictor is better than LES, Avg-LES, P-LES, and other DH predictors [15]. For each queue $j$, we store for each queue length $q \in \{1, \ldots, Q\}$, the waiting times of the last $N_q$ (a fixed integer greater than 0) customers that observe a queue equal to $q$ at their arrival. So for a new customer of type $j$ who observes a queue length $q$, the average of these $N_q$ waiting times will be the prediction of his waiting time.

Unlike Avg-LES, here a predictor for which $N_q$ is large performs better than a predictor with $N_q = 1$. For AvgC-LES to be efficient, $N_q$ must be large, and for this a lot of memory is needed to store the $N_q$ delay for all $q \in \{1, \ldots, Q\}$. However in most call center software, there is not enough memory to store all this information. This is the main drawback of this predictor and this is one of the reasons why we propose a new version that uses less memory and gives even better performance measures.

### 4.6 Weighted-Average-LES-Conditional-on-Queue-Length (WAvgC-LES)

To solve the need for a large quantity of memory with AvgC-LES, we propose a new predictor which is an "average versions" of AvgC-LES, which replace the ordinary average of the $N_q$ wait times for class $j$ and queue size $q$ by a "weighted average". In this work, we use exponential decreasing weighting. To ensure that each new observation (each waiting time) makes a relatively small contribution, we have chosen small $\alpha$ smoothing factors (e.g., 0.2 or smaller). With exponential smoothing, we no longer need to store any individual information on customer waiting times. In our numerical experiments, WAvgC-LES gave better or similar results than AvgC-LES in terms of prediction error.

Here we describe how to predict the delay for a new arrival customer in system who find $q$ customer in queue. When the arrival customer is the first one who observe this queue length (so we have $S_q = -1$), so his wait time is predicted by LES, otherwise his wait time is predicted by $S_q$ value.

Now we describe how to update $S_q$. When a customer, who have found $q$ customers in queue at its arrival and after waited a delay $W$, exits the queue to receive service, the value of $S_q$ is update by

$$S_q = W \tag{1}$$

if its value is minus one, else it is updated by an exponential smoothing average

$$S_q = \alpha \cdot W + (1 - \alpha) \cdot S_q \tag{2}$$

where $\alpha$ is the smoothing factor, and $0 < \alpha \leq 1$.

Based in our numerical experiment, we observe that in system with time varying arrival process and time varying servers as the modern call centers, a small value for the smoothing factor give better predictions than large value. This mean that it is better to give small weight to new LES and large weight to the old LES values for this system. We recommend $\alpha = 0.2$. We observe also that in single queue system with long run simulation, the precision of WAvgC-LES is very similar to that of QL. This can be explained by the fact that WAvgC-LES has gathered sufficient data to calculate a good expected waiting time conditional on queue length.

## 5 Numerical Model Results

In this section, we present the numerical results of experiments with simulated models, as well as the numerical results of experiments with a real multiskill call center. We compare the precision of the predictors on all the models studied. We start with the M/M/$s$+M model, for which we have an analytical formula for calculating the expectation of waiting time conditional on queue length. This example is studied in order to check the accuracy of our predictions in relation to the actual value. Our second example is an N (multi-skilled) model (see Fig. 1). It has two types of customer and two groups of agents. Agents in group 1 have the skill to serve only type 1 customers, and agents in group 2 can serve any type of customer. Our third and final example is an actual multi-skilled call center. It has several call types (27 in all) and several agent groups.

### 5.1 Measure of the Prediction Errors

To measure the accuracy of our delay predictors, we use the *"Mean Squared Error"* (MSE). Let $D$ be the predicted waiting time of a customer and $W$ the actual waiting time. The MSE is given by

$$\text{MSE} = \mathbb{E}\left[(W - D)^2\right].$$

In practice, we use its empirical version, called the *"Average Squared Error"* (ASE), The ASE is given by

$$\text{ASE} = \frac{1}{N} \sum_{n=1}^{N} (W_n - D_n)^2,$$

where $D_n$, $W_n$, and $N$ are respectively the predicted waiting time, the real waiting, and the total number of served customers. The normalized version of the ASE called *"Root Relative Average Squared Error"* (RRASE) is reported in numerical results.

$$\text{RRASE} = \frac{\sqrt{\text{ASE}}}{\sum_{n=1}^{N} W_n/N} \times 100.$$

### 5.2 An M/M/$s$+M Model Queue System

We compare the accuracy of the predictors in an M/M/$s$+M model. We consider that the arrival rate varies over the day. The day is divided into 20 periods of 1 h each, and in each period $p$ arrivals follow a Poisson process of constant rate $\lambda_p$. Service time follow an exponential distribution with rate 1. The distribution of patience times is also exponential with rate 0.5. We take $s = 20$ for the whole day, $\lambda_p = 25$ for even-numbered periods, and $\lambda_p = 20$ for odd-numbered periods. We run 100 independent simulations and estimated the accuracy of the delay predictors. We observe that the average waiting time is equals to 20 min, the average queue length is 8 customers, the probability of a customer waiting is around 92%, and the probability that a customer leave the queue without served is 15.8. For this model, it is well known that QL is an optimal predictor of waiting time expectation conditional on queue length. Our objective is to compare the performance of the other predictors with QL predictor to see how they score against the optimum.

Table 1 shows the RRASE values for the different predictors. Here are the predictors parameters used to compute the RRASE. $N_j = 2$, $N_{j,k} = 100$, and $\delta = 0.1$ respectively for Avg-LES, AvgC-LES, and WAvgC-LES. QL gives the best performance, which is no surprise. It's closely followed by AvgC-LES and WAvgC-LES, which give roughly the same result. The other methods perform much less well. They give much larger RRASEs. It should be noted that Avg-LES ($N_j \geq 2$), which is often used in practice, performs less well than LES ($N_j = 1$). Ibrahim et al. [6] found comparable results. P-LES is the worst predictor.

**Table 1.** The RRASE result for the M/M/20+M simulated example.

| | LES | Avg-LES | P-LES | E-LES | AvgC-LES | WAvgC-LES | QL |
|---|---|---|---|---|---|---|---|
| RRASE | 46.9 | 49.4 | 59.2 | 43.6 | 32.9 | 32.8 | 32.1 |

### 5.3 The N-Model Example

An N-model is illustrated in Fig. 1 by the middle image. The routing policy used is as follows. Calls are served according to their order of arrival in each queue (FCFS order). Agents of group 2 give priority to calls type 2. They serve calls

type 1 only if there are no type 2 calls waiting. If a call type 1 arrives, an agent of group 1 who has been idle for the longest time is preferred. If no agent of group 1 is free, the call is routed to the Group 2 agent who has been inactive the longest. If all agents in group 2 are busy, the call is placed in queue 1.

We divide the day into 10 periods of 1hour, and for each period arrival follow a Poisson process with constant rate. The service times and patience times are exponential with constant rates over the day. The parameters for model simulation are $\lambda_1 = (25, 34, 43, 48, 51, 57, 42, 34, 22, 18)$ per hour, the vector of arrival for type 1, $\lambda_2 = (26, 40, 47, 59, 68, 59, 48, 43, 39, 29)$ the vector of arrival rate for call type 2. The mean service time for call type 1 is $\mu_1^{-1} = 21$ minutes and their mean patience is $\nu_1^{-1} = 46.7$ minutes. For call type 2, the mean service time is $\mu_2^{-1} = 11$, and mean patience time is$\nu_2^{-1} = 30$. The staffing vectors are $s_1 = (4, 6, 9, 10, 9, 9, 9, 8, 5, 5)$ for group 1 and $s_2 = (4, 7, 9, 10, 9, 8, 7, 8, 6, 5)$ for group 2. We run 100 independent simulation of days. We observe that only 22% of calls type 2 are served by group 2, and the 88% by group 1. The probability of delay is 94.0 % for call type 1, and 97% for call type 2. The ratio of abandonment is 33% and 23% for call type 1 and call type 2 respectively. The average queue length is 9.7 for type 1 and 5.5 for type2. The average waiting for call type 1 is 938 s and 426 s for call type 2.

Table 2 shows the RRASE for both types of call, for different predictors. We use $N_j = 7$, $N_{j,k} = 100$, and $\delta = 0.2$ for Avg-LES, AvgC-LES, and WAvgC-LES respectively. The QL predictor is not usable in a multi-skill context, so here we use ANN predictor, which is known to perform best in a multi-skill context [13]. Unsurprisingly, ANN, which needs a learning steep, a lot of data, and difficult to implement in practice, gives the best results. After the ANN, WAvgC-LES is DH predictors who gives the best result. P-LES, as in the other examples, still gives the worst performance. LES, Avg-LES and E-LES gives close results.

**Table 2.** The RRASE of the N-model.

| Call Type | LES | Avg-LES | P-LES | E-LES | AvgC-LES | WAvgC-LES | ANN |
|---|---|---|---|---|---|---|---|
| T1 | 49.5 | 51.8 | 70.4 | 46.8 | 37.5 | 36.4 | 32.4 |
| T2 | 62.1 | 66.7 | 94.3 | 61.4 | 47.8 | 44.3 | 41.2 |

### 5.4 The Real System Example

The real system studied in this work is a multi-skilled call center situated in the Netherlands. There are two datasets collected during the year 2014. The first dataset concerns call information logs and the second is a dataset on the different activities of agents during the day at the call center. The call log data set contains information on arrival time of call, the begin and the end service time of a call, the type of service asked by a caller, information that identify the

agent, etc. The activity data contains the identity of the activity, the begin and end time of an activity, etc.

The call center is open 12 h a day. It opens at 8 a.m. and closes at 8 p.m. Monday to Friday. There are 27 possible types of service, and 312 separate agents worked over the year. An analysis of the data showed 56% of calls are answered immediately with no waiting time, 38% of callers had to wait in a queue before receiving service, and around 6% of customers leave queue before receiving service. In this work, we report only the results of the 5 call types (T1, T2, T3, T4, and T5) that received nearly 90% of call volume (Table 3).

**Table 3.** Some statistical summary over the year for the real system.

| | T1 | T2 | T3 | T4 | T5 |
|---|---|---|---|---|---|
| Total number calls | 568 554 | 270 675 | 311 523 | 112 711 | 25 839 |
| Served, no wait | 61% | 52% | 55% | 45% | 34% |
| Served, waited | 35% | 40% | 40% | 46% | 54% |
| Abandon | 4% | 7% | 5% | 8% | 12% |
| Avg wait time (sec) | 77 | 91 | 83 | 85 | 110 |
| Avg service time (sec) | 350 | 308 | 281 | 411 | 311 |
| Avg queue length | 8.2 | 3.3 | 4.4 | 4.3 | 0.9 |

Table 4 shows the RRASEs for different predictors in the real system for the five that have received bigger call type volume. We use $N_j = 10$, $N_{j,k} = 100$, and $\delta = 0.2$ for Avg-LES, AvgC-LES, and WAvgC-LES respectively. As in the previous example, we compare DH predictors with ANN predictor from Thiongane et al. [13,15]. ANN predictors are more efficient, but they require a very costly learning phase and involve many parameters. In this example also we also observe that our new WAvgC-LES predictor is the better DH predictor, far behind ANN. WAvgC-LES is shortly followed by AvgC-LES. As observed in Ibrahim et al. [3] LES is better than Avg-LES. As in two other example, we notice that the performance of P-LES is always bad.

**Table 4.** RRASE for the five call types of real system.

| Call Types | Delay Predictors | | | | | | |
|---|---|---|---|---|---|---|---|
| | P-LES | Avg-LES | LES | E-LES | AvgC-LES | WAvgC-LES | ANN |
| T1 | 81.75 | 76.28 | 58.66 | 68.59 | 56.97 | 56.62 | 42.24 |
| T2 | 98.01 | 74.09 | 61.02 | 56.27 | 60.87 | 58.49 | 44.18 |
| T3 | 94.87 | 82.19 | 62.43 | 67.86 | 63.68 | 62.44 | 48.32 |
| T4 | 92.96 | 82.22 | 63.53 | 69.68 | 63.12 | 62.17 | 50.25 |
| T5 | 92.44 | 70.87 | 53.47 | 53.54 | 53.20 | 51.28 | 39.47 |

## 6 Conclusion

In this work, we develop and compare a delay history predictor for multi-skill call centers. The new delay history predictor compute an exponential smoothing average of wait times of the past customer who observe the same queue length at their arrival. We find that our new DH predictor performs much better than other DH predictors. The ANN predictor, which is difficult to implement in practice, is better than the WAvgC-LES, but the latter is easy to use in practice and also gives fairly accurate performance. In this work, the predictions are point estimate of the waiting time. The prediction is an estimate of the expected waiting time conditional on the queue length and other system parameters when the customer enters the queue. As part of our ongoing work, we aim to develop efficient methods for predicting and announcing the expected waiting time and its variance that are conditional on the system state when a customer enters in queue.

**Acknowledgements.** Thanks to Ger Koole (VU Amsterdam) for the data provided.

## References

1. Ang, E., Kwasnick, S., Bayati, M., Plambeck, E., Aratow, M.: Accurate emergency department wait time prediction. Manuf. Serv. Oper. Manag. **18**(1), 141–156 (2016)
2. Armony, M., Shimkin, N., Whitt, W.: The impact of delay announcements in many-server queues with abandonments. Oper. Res. **57**, 66–81 (2009)
3. Dong, J., Yom Tov, E., Yom Tov, G.: The impact of delay announcements on hospital network coordination and waiting times (2016)
4. Gans, N., Koole, G., Mandelbaum, A.: Telephone call centers: tutorial, review, and research prospects. Manuf. Serv. Oper. Manag. **5**, 79–141 (2003)
5. Garnett, O., Mandelbaum, A., Reiman, M.: Designing a call center with impatient customers. Manuf. Serv. Oper. Manag. **4**(3), 208–227 (2002)
6. Ibrahim, R., Armony, M., Bassamboo, A.: Does the past predict the future? The case of delay announcements in service systems. Manag. Sci. (2016)
7. Ibrahim, R., Whitt, W.: Real-time delay estimation based on delay history. Manuf. Serv. Oper. Manag. **11**, 397–415 (2009)
8. Ibrahim, R., Whitt, W.: Real-time delay estimation in overloaded multiserver queues with abandonments. Manage. Sci. **55**(10), 1729–1742 (2009)
9. Ibrahim, R., Whitt, W.: Real-time delay estimation based on delay history in many-server service systems with time-varying arrivals. Prod. Oper. Manag. **20**(5), 654–667 (2011)
10. Senderovich, A., Weidlich, M., Gal, A., Mandelbaum, A.: Queue mining – predicting delays in service processes. In: Jarke, M., Mylopoulos, J., Quix, C., Rolland, C., Manolopoulos, Y., Mouratidis, H., Horkoff, J. (eds.) CAiSE 2014. LNCS, vol. 8484, pp. 42–57. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-07881-6_4
11. Senderovich, A., Weidlich, M., Gal, A., Mandelbaum, A.: Queue mining for delay prediction in multi-class service processes. Inf. Syst. **53**, 278–295 (2015). http://dx.doi.org/10.1016/j.is.2015.03.010

12. Senderovich, A., Weidlich, M., Gal, A., Mandelbaum, A.: Queue mining for delay prediction in multi-class service processes. Inf. Syst. **53**, 278–295 (2015)
13. Thiongane, M., Chan, W., L'Ecuyer, P.: Waiting time predictors for multiskill call centers. In: Proceedings of the 2015 Winter Simulation Conference, pp. 3073–3084. IEEE Press (2015)
14. Thiongane, M., Chan, W., L'Ecuyer, P.: New history-based delay predictors for service systems. In: Proceedings of the 2016 Winter Simulation Conference, pp. 425–436. IEEE Press (2016)
15. Thiongane, M., Chan, W., L'Ecuyer, P.: Delay predictors in multi-skill call centers: An empirical comparison with real data. In: Proceedings of the International Conference on Operations Research and Enterprise Systems (ICORES), pp. 100–108. SciTePress (2020)
16. Thiongane, M., Chan, W., L'Ecuyer, P.: Learning-based prediction of conditional wait time distributions in multiskill call centers. In: Parlier, G.H., Liberatore, F., Demange, M. (eds.) Operations Research and Enterprise Systems, pp. 83–106. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-10725-2_5
17. Whitt, W.: Predicting queueing delays. Manag. Sci. **45**(6), 870–888 (1999)