The Impact of Agents Heterogeneous in Call Center Performance Measures

Mamadou Thiongane(✉), Mohamed M. Ould Deye, Modou Gueye, and Ndiouma Bame

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

Abstract. Modern call centers are highly complex queuing systems, in which there are several possible call types, and many agents groups. Due to this complexity, simulators are now preferred for their management than standard Erlang queuing models. Several studies have shown through data analysis that agents often have quite different speeds for processing a call type. However, this agent heterogeneity is often neglected, which is why most simulators have not been designed to use heterogeneous agents to process a service. In this work, we use simulation to study the impact of agent heterogeneity on the performance of call centers. We developed and integrated a module into the call center simulator, *ContactCenter*, which allows the use of heterogeneous agents to process a service. We have analyzed data collected from a real call center and have shown that agents have different processing speeds for handling a call type. We have modeled the distribution of arrivals, service times and patience times that will enable us to simulate this call center. Simulation result show that call center performance on a given day depends largely on the agents chosen to answer the calls.

Keywords: Agents Heterogeneous · Call Center · Real Data

1 Introduction

A call center is a set of resources for communication between an organization and its customers over the phone. Today, we also refer to them as *contact centers*, because agents can use other mediums of communication, such as post, e-mail, online chat, etc. Call center is widely used in various service and manufacturing industries. Effective call center management is a difficult task because of the considerable sources of uncertainty; these include call arrival rates, which typically vary over time and are stochastic, service times, which are random and whose distribution may depend on the type of call and the agent handling it, as well as agents who may be absent or may not follow their planned schedules [3,6,7,10,11,14]. Given to the complexity of modern call centers, simulators are much more widely used for call center management than standard Erlang models.

K. Cheikh M. F. Kebe et al. (Eds.): InterSol 2024, LNICST 610, pp. 228–239, 2025.
https://doi.org/10.1007/978-3-031-86493-3_18

In this paper, we focus on the impact of agent heterogeneity on call center performance measures. We analyze data gathered at the call center of an information technology (IT) company in the Netherlands. This real call center setting is complex, consisting of many heterogeneous agents and multiple distinct call types. The data show that service times differ greatly across such agents. However, in call center management, this heterogeneity is often neglected, and it is often assumed that the distribution of call duration depends only on the type of call but not on the agent handling the call. That's why most of the simulators tools, such as *ContactCenters* [5], are not designed to take into account agent heterogeneity. They only allow the use of a single distribution of service time for a call type. To be more precise, the same distribution is used to generate service time for all agents that can serve a call type.

In this work, we will develop and add a module in *ContactCenters* that allows the specification of a distribution of service times for each pair (agent, call type). *ContactCenters* is a call center simulator developed with Java by the Stochastic Simulation Laboratory at the University of Montreal (Canada). It is also used by some companies to manage their call centers. We use this new version of the simulator and show that a simulation model that takes into account agent heterogeneity predicts call center performance better than a simulation model that ignores agent heterogeneity. However, before conducting the simulations, as a first step we will search through real data the distributions that best fit arrivals, service times, and patience times.

The remainder of this paper is structured as follows. Section 2 presents a literature review on agent heterogeneity in service systems, in particular for call centers. In Sect. 3, we describe and do a preliminary analysis of the data set that motivated this research. Section 4 present our call center modeling parameters, and the simulation experiments we conduct to show the impact of agent heterogeneity on call center performance. The conclusion is given in Sect. 5.

2 Litterature Review

To analyze call center operations, standard Erlang queueing models have been widely used. In these models, arrival are modeled as Poisson process, agent service times are modeled as independent, identically distributed exponential random variables with a constant mean. However, many studies have shown that the lognormal distribution is a remarkably good fit for the service-time distribution than the exponential distribution [4,7,11,15]. This is inevitably affecting call center performances. In queueing models, customer heterogeneity has received ample attention in both practice and theory. In contrast, server heterogeneity has received relatively scarce attention. There are not much research addressed on the statistical and practical implications of service time heterogeneity among agents. Some works on routing policies, which studied queueing models with heterogeneous servers, try to route incoming calls to minimize a performance measure, such as the average waiting time. Most of them try to find the optimal routing policies in large-scale systems under heavy-traffic conditions [1,2,8,9].

In general, these papers show that control decisions can actually benefit from agent heterogeneity, e.g., routing incoming calls to the fastest idle agents reduces customer waiting. Mehrotra et al. [13] do a numerical study to characterize overall performance in terms of customer waiting time and overall resolution rate. Wang et al. [16] study scheduling and routing strategies of heterogeneous agents in call centers. They construct an integer linear programming of the scheduling problem for call centers with agent heterogeneity, and combine the use of a discrete-event simulation model with an artificial bee colony algorithm to solve the model.

There is very little empirical research supporting that theoretical work. Gans et al. [7] analyze call-center data and identified both short-term and long-term factors associated with agent heterogeneity in practice. Ibrahim et al. [11] use mean service time from real data and propose a method to predict the variance of service times. Assuming that service times follow a lognormal distribution (that uses the mean and the predicted variance), they show through small simulation models that agents heterogeneity can have an impact on call center performance measures.

In this paper, we extend the theoretical research mentioned above with empirical work. We have analyzed data and shown that service times differ greatly across agents. We show through our data that the log-normal distribution fits service durations better than the exponential distribution, the call arrival follow a non-homogeneous Poisson process, and customers patience follow an exponential distribution. We will show through simulation that call center performance measures could be much closer to real performance when the agents heterogeneity is taken into account. Thus, we take a step forward to fill this gap in the literature.

3 Data and Analysis

In this section, we will describe the two datasets collected in the call center studied, and we will also perform an in-depth analysis of these data.

3.1 Datasets Description

In this work, we use two dataset collected by VANAD Laboratories located in Rotterdam, in The Netherlands. They were collected over the span of one year, ranging from January 1, 2014 to December 31, 2014. This center operates from 8 a.m to 8 p.m from Monday to Friday. Unlike most call center data, which are only aggregated data (that are not always complete to extract a day's parameters and performance measures), here we have call-by-call log data and agents activities data. In our data set, there are 27 call types. We call them T1, T2, $\cdots$, T27 from the one with the highest call volume to the one with the lowest. They are handled by a group of 312 agents. This includes part-time agents, full-time agents, agents that worked only for a few months and agents that worked in every month of the year. Each agent has a skill set, which consists of at least one skill. Not every agent has all the skills. In total, there are 2,983 distinct

agent/call type combinations, where each combination corresponds to an agent handling a particular call type. Our data contains a total of 1,543,164 call logs and 1,639,770 activities logs.

The call-by-call data contains on each received call the following information: the call type, the arrival time, the date of the day, the desired service, the Voice Response Unit (VRU) entry and exit time. For the calls that have to wait, the data contains the queue entry time and the queue exit time. For each received call, we know whether it has been served or abandoned. When a call is abandoned, the time of abandonment is also known. Finally, for a served call, we have the started service time, the ended service time, and the ID of the agent who serve the call.

Activity data contains information on the activities carried out by an agent during a working day at the call center. This information includes the activity ID, the activity start time, the activity end time, and the agent ID.

3.2 Data Analysis

To sketch a temporal distribution of the workforce, we plot in Fig. 1 the average number of agents answering calls per weekday, with 95% confidence bands. We see that the number of agents is highly variable on Mondays, and that Fridays have the least number of agents, on average. In Fig. 2, we plot the total average call volume per weekday, including all call types. Consistent with Fig. 1, Fig. 2 shows that call volumes on Mondays exhibit the highest variance, and that call volumes on Fridays are lowest on average.

Figure 3 gives a scatter plot of the empirical means versus variances of service times for different call types in our data. Each point in the plot corresponds to a given (mean, variance) pair, corresponding to a given call type. Figure 3 shows that there are significant differences in means and variances across different call types. As expected, Fig. 3 shows that call types with longer durations generally exhibit higher variances.

Service time distributions for the same call type vary considerably depending on the agent. In Figs. 4 and 5, we illustrate this agent heterogeneity. We plot average service times for two call types: $T1$, which is handled by 286 agents, and $T2$, which is handled by 191 agents, as a function of the total number of calls answered (over the one-year period covered by our data) by each agent.

The horizontal line in each figure indicates the overall average service time across all agents, for each call type. Figures 4 and 5 show that there is significant variability in service times across all agents. Figures 4 and 5 also show that there are clearly clusters of agents who seem to perform in a roughly similar manner (having either shorter or longer than average service times). In general, agents who have handled many calls during the year are much faster on average than those who have handled few calls. The latter are either agents who have handled very few calls in general, or ones who have mostly handled calls of other types. To illustrate this, in Fig. 6, we plot the average service time of each months of some experienced agents and some new agents. As one can see from Fig. 6(a) the average service time of each experienced agents is different; furthermore, in Fig. 6(b), the average service time of new agents all exhibit a declining trend,

which suggests that new agents learn over time, and their average service time decrease as they learn.

In Figs. 7 and 8, we plot the variances of service times for all agents handling call type $T1$ and $T2$, respectively, as a function of the total number of calls of that type answered by the agent. It appears that agents who have handled more calls tend to exhibit less variance in their service times. In other words, the larger dispersion is mainly exhibited by less experienced agents (those answering fewer calls). These Figures confirm that there are clear differences in variance of service times across agents.

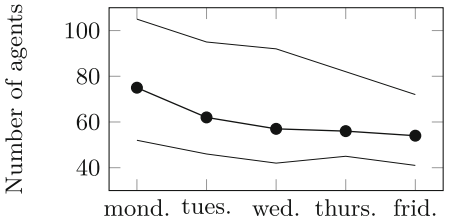

Fig. 1. Average number of agents per weekday and corresponding 95% confidence bands.

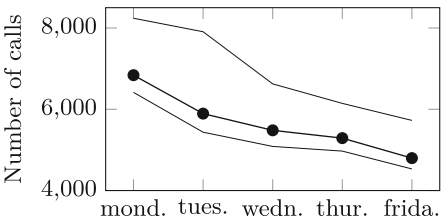

Fig. 2. Average number of calls per weekday and corresponding 95% confidence bands.

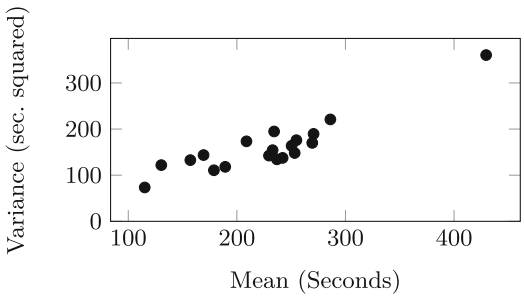

Fig. 3. Each point corresponds to a (mean, variance) pair for a given call type.

4 Call Center Modeling and Simulation Experiments

In this section, we present the call center parameters modeling, and we describe the simulation experiments that we conduct to evaluate the impact of agent heterogeneity on call center performances measures.

4.1 Call Center Modeling

As we said earlier, the call center studied in this work have $K = 27$ different call types. There is one waiting queue per call type. In this section, we describe how

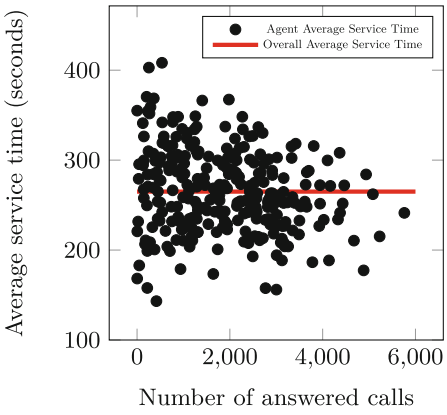

Fig. 4. Average service times for different agents handling type $T1$ calls as a function of the total number of calls answered per year.

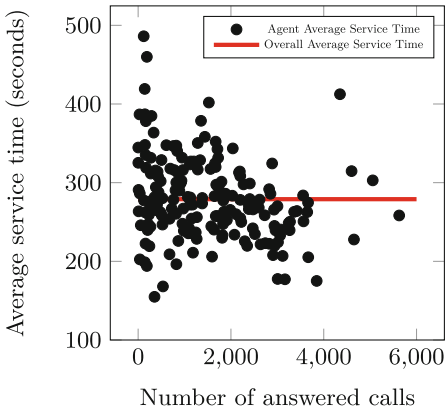

Fig. 5. Average service times for different agents handling type $T2$ calls as a function of the total number of calls answered per year.

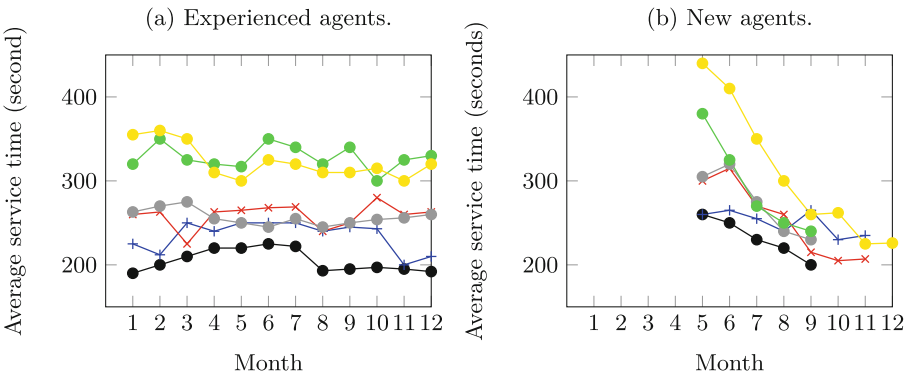

Fig. 6. Average service time per month of some agents.

we model call arrivals, service times, and patience times for use in a call center simulation. Note that in this section, we will report the results of modeling of one or two call types. However, it should be noted that the results are similar for all call types. It should also be noted that, prior to modeling, we removed the bad days from our dataset, i.e. holiday days and very special days. There are 21 such days in the year.

Arrival Process: Figure 9 and 10 shows the annual mean of arrival counts per period of 30 min and per weekday for call type T1 and call type T2, respectivrly. We see from these figures that the arrival behavior for Monday differs significantly from that of the other days and the arrival rate varies considerably during the day. Figure 11 shows the fit of call type T1 inter-arrival data with an exponential distribution. We observe that the exponential distribution fits the inter-arrival well, so we can deduce that arrivals follow a Poisson process. The

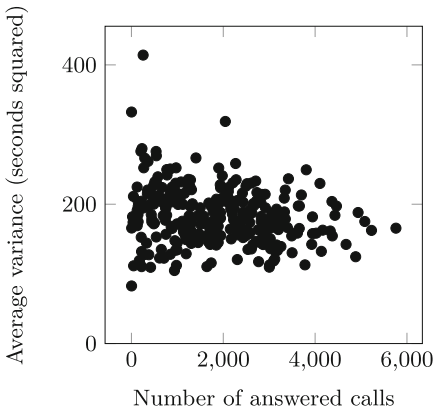

Fig. 7. Average variances of service times for agents handling type T1 calls as a function of the total number of calls answered per year.

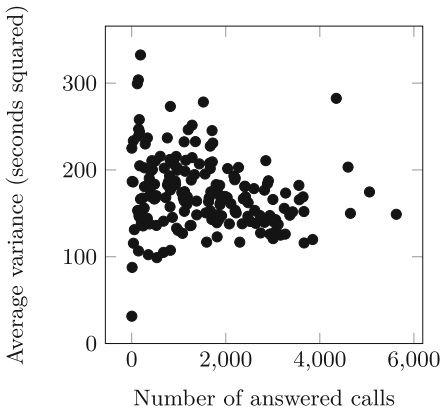

Fig. 8. Average variances of service times for agents handling type T2 calls as a function of the total number of calls answered per year.

result observed with Fig. 9 and Fig. 11 combined shows that the arrivals follow a non-stationary Poisson process. To take this into account in the simulation, we have divided the opening hours of the call center into $P = 24$ time periods of 30 min. So, for a call type k, the arrival process is a Poisson process with a constant rate $\lambda_{k,p}$ over each period p, so the vector of arrival rates over the P periods is $\lambda_k = (\lambda_{k,1}, \ldots, \lambda_{k,P})$.

Service Time: A service time often consists of a first part handled by an interactive voice response (IVR) system, and a second part where the call is handled by an agent. Since we are interested in service times from the viewpoint of agents, we do not consider the IVR part because agents are not involved for that part. From the viewpoint of an agent, an individual service time is the sum of: (i) the time spent actually talking to the customer (call time), and (ii) the post-call time spent by the agent to wrap up issues related to the call, during which s/he remains unavailable.

Figure 12 shows the fit of call type T1 service time data with an exponential and lognormal distribution. We observe that the lognormal fit better than the exponential distribution. In our simulation, for type k, the service times distribution are modeled by a lognormal distribution with mean μ_k and variance σ_k^2. Notice that to take into account the heterogeneity of agents we specify a distribution of service times for each pair (agent, call type).

Patience Time: The patience time represent the time a customer is willing to wait for service. A customer abandons the queue once her waiting time exceeds her patience time. Figure 13 shows the fit of call type T1 patience time data with the exponential distribution. We observe that the exponential distribution fits the patience time well. In simulation, the patience times for a call type k are exponential with mean $1/\nu_k$.

The Staffing: For each working day agents are divided into G groups. An agent of group $g \in \{1, \ldots, G\}$ has the skill set $\mathcal{S}_g \subseteq \{1, \ldots, K\}$ which defines the set of call types she can serve. Let $s_g = (s_{g,1}, \ldots, s_{g,P})$ be the staffing vector of group g, where $s_{g,p}$ is the number of agents from that group working in period p. For each working day, G the number of group, and s_g the staffing vector of group g can be calculated from the log activity data.

Routing Policy: The routing mechanism works as follows. When a customer calls, she will interact with the IVR by making use of her key pad to choose the call type k. If there is any agent available with the skill to handle that type of calls, then she is routed to the longest idle agent of those available agents; otherwise, she will wait in an invisible queue. The calls in this queue are served in the FCFS (first come first served) order.

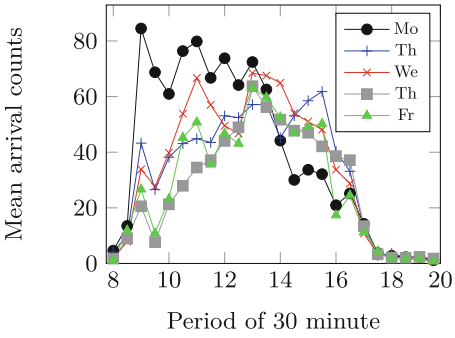

Fig. 9. Annual mean of arrival counts per 30 min and per weekday for call type T1

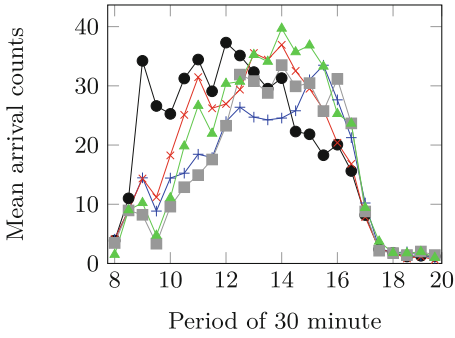

Fig. 10. Annual mean of arrival counts per 30 min and per weekday for call type T2.

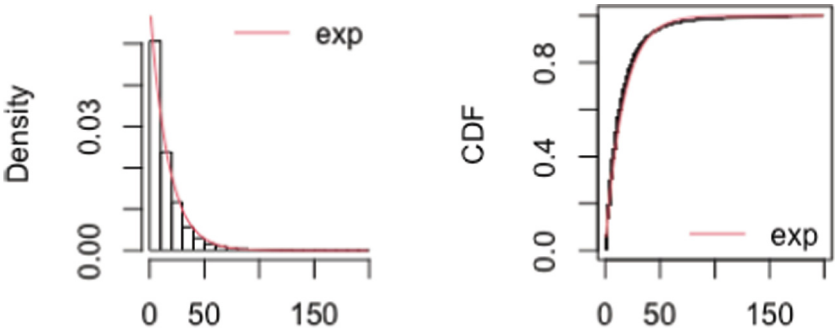

Fig. 11. Fit inter arrival with Exponential distribution for call type T1.

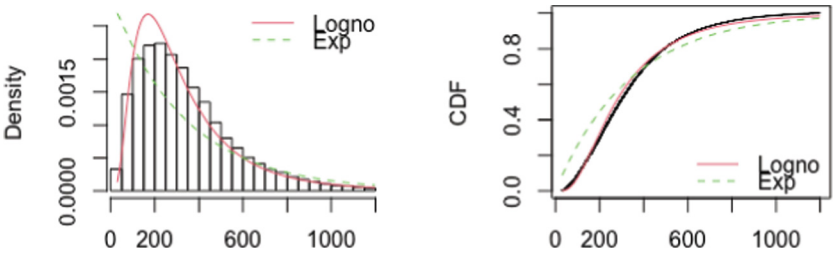

Fig. 12. Fit service time distribution for agent $a1$ for type T1 with Exponential and Lognormal.

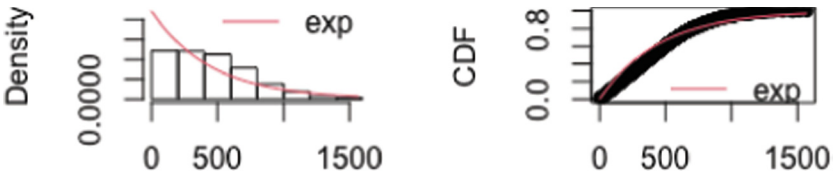

Fig. 13. Fit patience time with Exponential distribution for call type T1.

4.2 Simulation Experiments

In this section, we compare the performance of two simulation models of the call center with actual call center performance. This comparison is made on two different days, Day 1 and Day 2 chosen at random in our dataset. The first is a day in May and the second a day in July. Here's a description of each model.

Model 1: the arrival process is a piecewise constant Poisson process for each call type k. The vector λ_k that represent the arrival rate over all period and $1/\nu_k$ the mean patience time for call type k are calculated from the real data of the simulate day; The used agents are those who worked on the simulated day. Indeed, the number of group G, and the staffing vector s_g for each group g are calculated from the data; the service time distribution is lognormal and we take into account the heterogeneity of the agents. Thus for each pair (agent a, call type k) the parameters $(\mu_{a,k}, \sigma_{a,k})$ of lognormal distribution are also calculated from the data; The routing policy is that described in the previous section.

Model 2: It is similar to Model 1 except on service time distribution. Here, we assume that all agents are identical, so the distribution of service times depends on the type of call and not on the agent handling the call. Here we have a distribution of service times for each call type. The parameters of the distribution of the service times for each call type are calculated using data. The mean (the variance) is average of the mean (the variance) service time of all agents who have the skill to handle that call type.

The performances we consider are the average waiting time (AWT) of calls and the service level (SL(s)), defined as the percentage of calls whose waiting times are less than s seconds (here we fixed $s = 60$). We are only interested in the performance of the 8 call types that receive more than 99% of call volume.

The system we simulate is non-stationary. Consequently, the SL and AWT results are different for each simulation. Given this variability, we repeat the simulation $n = 10,000$ times. Therefore, for each model and for each day, there are 10,000 simulation results and one realization. To measure the difference between simulated and actual results, we use WAE (weighted absolute errors).

$$\mathrm{WAE_X} = \frac{\sum_{i=1}^{n} \mathrm{A}_i \mid \mathrm{E}\hat{\mathrm{X}}_i^{sim} - \mathrm{X}_i^{act} \mid}{\sum_{i=1}^{n} \mathrm{A}_i},$$

where $\hat{\mathrm{X}}_i^{sim}$ is the simulated result of day i, and X_i^{act} is the actual result of day i, and A_i is the number of arrivals in day i. We compute the WAE of SL and AWT, which is $\mathrm{WAE_{SL}}$, $\mathrm{WAE_{AWT}}$, respectively.

WAE measures the difference between the simulation results and the actuals. If WAE is equals to 0 this means that the simulation model has produced a result that is equal to the actuals. In other words, the model is a perfect representation of the system to be simulated. However, this ideal does not exist in stochastic modeling, and WAE are always positive. One part of the WAE comes from the variability in SL, and AWT; for exemple the SL is different per simulation. The other part of WAE comes from the model; for example, if one simulates a model which does not describe the reality well, then there is a big difference between the simulation results and the actuals.

4.3 Results and Discussion

In Fig. 14(a) and Fig. 14(b), we plot the $\mathrm{WAE_{AWT}}$ values for Model 1 and 2 as function of the call type, for Day 1 and Day 2, respectively. We observe that the Model 1, which takes into account agent heterogeneity, have a $\mathrm{WAE_{AWT}}$ that is around 20 seconds less than that of the Model 2, which assumes that all agents are identical, for all call types. This means that Model 1 predicts the average waiting time better than Model 2. Figure 14(c) and Fig. 14(d) plot the $\mathrm{WAE_{SL}}$ for Model 1 and 2 as function of the call type, for Day 1 and Day 2, respectively. We can see that the Model 1 have a $\mathrm{WAE_{SL}}$ that is around 1.5% less than that of the Model 2. This means that Model 1 predicts the service level better than Model 2. With these results, we can conclude that Model 1 models the call center better than Model 2.

The difference in performance between the two models may appear minimal at first glance, however they could lead to significant cost savings in practice. ACS Wireless found that by reducing AWT by just 0.6 s will save $8 million a year [7,12]. In addition, small percentage differences in SL can make the difference between compliance and breach of service level agreements, which can lead to heavy penalties for the call center [7,11].

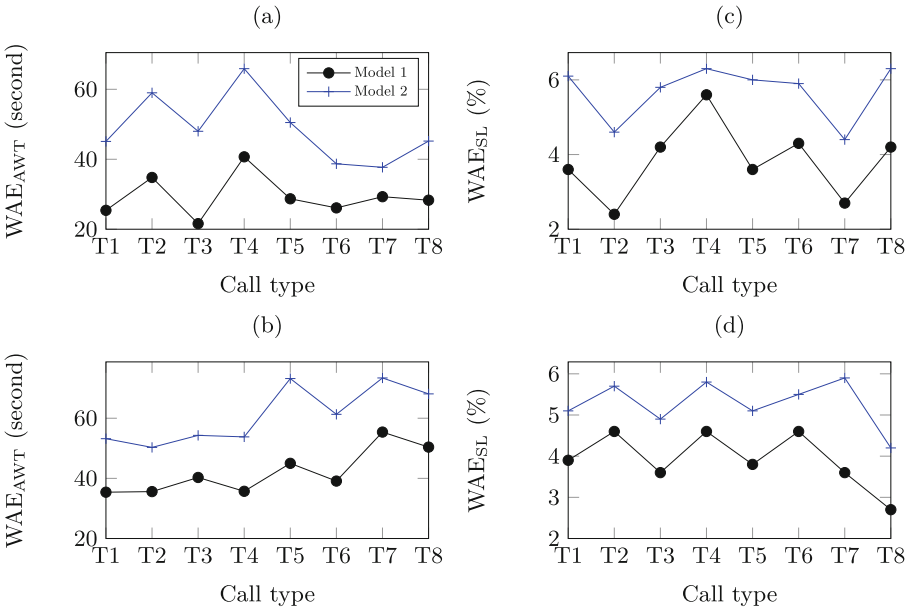

Fig. 14. WAE for Model 1, and Model 2 on Day 1 and Day 2

5 Conclusion

In this work, we looked at the impact of agent heterogeneity on the performance measures of a call center in which we have detailed data. We have reviewed the literature on agent heterogeneity in call centers. We analyzed real data and showed that agent service times vary considerably depending on the agent for a given service. A module for simulating a call center with heterogeneous agents was developed and added to the *ContactCenter* simulator. Modeling and parameter estimation with real data for the distribution of arrivals, service times, and patience times was done to enable us to simulate the call center. By simulating two call center models on two separate days, we have shown that taking into account agent heterogeneity leads to better predictions of call center performance. In future work, we plan to study the impact of unplanned breaks and correlation on the service times of call types handled by the same agent.

Acknowledgements. We thank Ger Koole, from VU Amsterdam, who provided the data.

References

1. Armony, M.: Dynamic routing in large-scale service systems with heterogeneous servers. Queue. Syst. **51**(3–4), 287–329 (2005)
2. Armony, M., Ward, A.R.: Fair dynamic routing in large-scale heterogeneous-server systems. Oper. Res. **58**(3), 624–637 (2010)
3. Avramidis, A.N., Deslauriers, A., L'Ecuyer, P.: Modeling daily arrivals to a telephone call center. Manag. Sci. **50**(7), 896–908 (2004)
4. Brown, L., et al.: Statistical analysis of a telephone call center: a queueing-science perspective. J. Am. Stat. Assoc. **100**, 36–50 (2005)
5. Buist, E., L'Ecuyer, P.: A Java library for simulating contact centers. In: Kuhl, M.E., Steiger, N.M., Armstrong, F.B., Joines, J.A. (eds.) Proceedings of the 2005 Winter Simulation Conference, pp. 556–565. IEEE Press (2005)
6. Gans, N., Koole, G., Mandelbaum, A.: Telephone call centers: tutorial, review, and research prospects. Manuf. Serv. Oper. Manag. **5**, 79–141 (2003)
7. Gans, N., Liu, N., Mandelbaum, A., Shen, H., Ye, H.: Service times in call centers: agent heterogeneity and learning with some operational consequences. In: Berger, J., Cai, T., Johnstone, I. (eds.) Borrowing Strength: Theory Powering Applications – A Festschrift for Lawrence D. Brown, vol. 6, pp. 99–123. Institute of Mathematical Statistics (2010)
8. Gurvich, I., Armony, M., Mandelbaum, A.: Service-level differentiation in call centers with fully flexible servers. Manag. Sci. **54**(2), 279–294 (2008)
9. Gurvich, I., Whitt, W.: Queue-and-idleness-ratio controls in many-server service systems. Math. Oper. Res. **34**(2), 363–396 (2009)
10. Ibrahim, R., L'Ecuyer, P., Régnard, N., Shen, H.: Modeling and prediction of service times in call centers (2013)
11. Ibrahim, R., L'Ecuyer, P., Shen, H., Thiongane, M.: Inter-dependent, heterogeneous, and time-varying service-time distributions in call centers. Eur. J. Oper. Res. **250**, 480–492 (2016)
12. J, H.: How the right headset affects call center productivity and the bottom line (2014). http://telecom.hellodirect.com/docs/tutorials/productivity.1.080701.asp. Accessed July 2015
13. Mehrotra, V., Ross, K., Ryder, G., Zhou, Y.P.: Routing to manage resolution and waiting time in call centers with heterogeneous servers. Manuf. Serv. Oper. Manag. **14**(1), 66–88 (2012)
14. Oreshkin, B., Régnard, N., L'Ecuyer, P.: Rate-based daily arrival process models with application to call centers. Oper. Res. **64**(2), 510–527 (2016)
15. Pichitlamken, J., Deslauriers, A., L'Ecuyer, P., Avramidis, A.N.: Modeling and simulation of a telephone call center. In: Proceedings of the 2003 Winter Simulation Conference, pp. 1805–1812. IEEE Press (2003)
16. Wang, M., Wang, X.: Study on the workforce scheduling and routing strategies of heterogeneous agents in call centers. In: Fifth International Conference on Economic and Business Management, vol. 159, pp. 577–583 (2020)