Performance Modeling for Call Centers Providing Online Mental Health Support

Tim Rens de Boer
CWI
Amsterdam, Netherlands
Email: trdb@cwi.nl

Saskia Mérelle
113 Suicide Prevention
Amsterdam, Netherlands
Email: s.merelle@113.nl

Sandjai Bhulai
Vrije Universiteit
Amsterdam, Netherlands
Email: s.bhulai@vu.nl

Rob van der Mei
CWI and Vrije Universiteit
Amsterdam, Netherlands
Email: mei@cwi.nl

Abstract—**Mental health helplines differ from other call centers, such as customer service call centers, in different aspects. Many of the agents handling conversations are volunteers, the conversations can be described as often more complex and/or emotional, and many of the mental health helplines use a triage system. More understanding is needed to improve staffing and/or service of mental health call helplines. Motivated by this, we propose a call center model that includes the specifics of online mental health helplines, including features such as a triage system for chats and the inclusion of service times consisting of warm-up, conversation, wrap-up and cool-down periods. This call center model is then validated using trace-driven simulation based on various months of real-life (anonymous) call and chat data provided by 113 Suicide Prevention. The model is validated by comparing the waiting times found in the data with the waiting times of the simulation. The results show that the model can simulate the waiting-time performance of the helpline accurately. Secondly, we focus on forecasting the number of arriving chats and telephone calls. Our results show that (Seasonal) Autoregressive Integrated Moving Average ((S)ARIMA) models trained on historical data perform better than other models in the case of short-term forecasting (five weeks or less ahead), while using linear regression works best for long-term forecasts (longer than five weeks). We propose a method on how these daily predictions can be quickly altered for hourly predictions, which can then be used together with the understanding of the model to obtain staffing advice.**

Keywords*—*call center models; queueing; mental health; helplines; data analytics; forecasting; staffing

I. Introduction

This paper is an extension of our previous research presented in [1], which was focused on the proposed call center model and validation. The present paper provides a more elaborate discussion on the analysis of real-life data, the validation of the model using trace-driven simulation, the error-term evaluation of demand forecasting, and how the model and forecasting can be used for staffing purposes.

Mental health helplines are helplines that are concerned with helping or assisting help seekers requiring mental health help, instead of physical care such as emergency helplines. There are many forms of mental health, with many countries having one or multiple helplines. Examples of mental health helplines in the Netherlands are 113 Suicide Prevention [2], the helpline for help-seekers with suicidal thoughts, the listen helpline (Dutch: luisterlijn) [3], and the Kindertelefoon [4] for children. Recently, mental-support helplines have received much attention due to the increase in call volumes related to the (partial) lockdown to combat the spreading of corona [5], [6]. This paper focuses on call center modeling for suicide prevention helplines. However, we emphasize that the results can also be used for modeling the waiting-time performance of other mental health helplines.

Suicide is a worldwide health problem. In 2020, on average, five persons died each day by suicide in the Netherlands alone. Suicide is a leading cause of death among adolescents [7]; worldwide, more than 700,000 people die from suicide every year [8]. In many countries, people struggling with suicidal thoughts can contact a helpline to get support to prevent and reduce suicidal thoughts [9]. In the Netherlands, persons with suicidal thoughts can contact 113, either by telephone or by chat [10], and are helped by volunteers and professionals. It is crucial that help seekers are answered swiftly. Therefore, it is important that adequate staffing is present to answer telephone calls and chat requests, minimizing the waiting times and abandonments (i.e., sudden termination of ongoing calls or chats while waiting). In order to calculate proper staffing levels (e.g., [11]), a good understanding of the processes of the call center system for mental health is required.

Mental health helplines differ in various aspects from classical call centers. First, the subjects of conversations are mental health concerns, e.g., loneliness and substance use [12]–[14]. Second, agents often have to handle complex conversations with the patient-in-need, and may themselves require support during or after a difficult conversation [17]. Therefore, an important aspect is that agents often need some time to cool down after emotionally difficult conversations. Third, when a chat or telephone call enters the system, it has to be determined which agent is best capable of handling the call, which results in a warm-up time before the call is taken into service. Lastly, chat conversations that enter the system first go through a triage phase. The inclusion of triage plays an important role, and functions as a filter [18] to chat requests, checking whether these help seekers are at the right helpline or might require emergency care. The triage employee can also estimate the conversation's difficulty to best match an agent with enough experience to handle the chat. So, the model for such a mental health helpline needs to meet the following requirements: (1) the possibility of abandonments, (2) a service time consisting of multiple phases, (3) a warm-up and cool-down time, (4) inclusion of chats and telephone calls, and (5) triage.

Classic call center models were considered for modeling this helpline, see [19] for an excellent overview of queueing models for call centers. The Erlang-C model is proven to be inaccurate when modeling call centers with abandonments [20]. The Erlang-A model includes abandonments [21], but does not include multiple skills, triage, and warm-up times. Multi-skill call center models, such as an N-configuration [19], were also

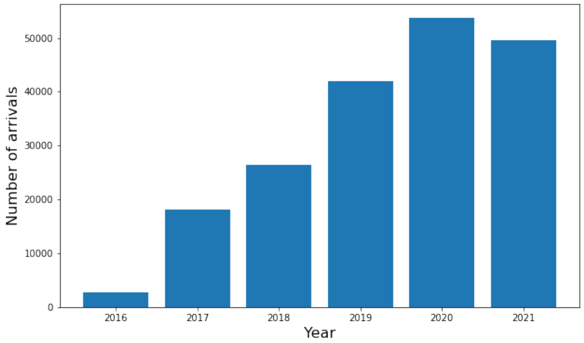

Figure. 1. The number of phone arrivals per year.

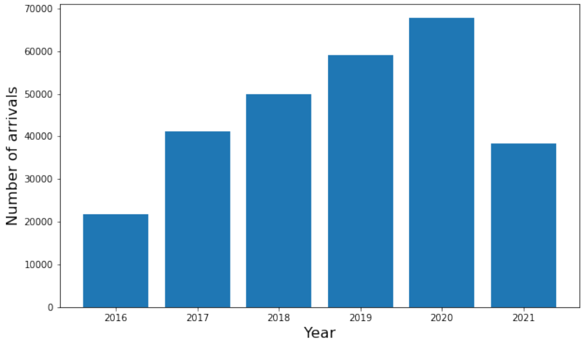

Figure. 2. The number of chat arrivals per year.

considered. The N-configuration model includes two types of arrivals, which in this case would be telephone and chat, and assumes that new chat arrivals can be picked up before triage. However, in the mental health helpline of [2], it is crucial that chats first go through triage. The importance of modeling triage is already shown in the emergency domain [23], [24]. Therefore, in this paper, we modify the N-configuration model in such a way that it includes the specifics of mental health helplines.

This paper introduces a new queuing model to include the possibility of triage. The model is built based on anonymized real-life call and chat data, made available by [2]. The records used cannot be traced back to individual callers, as only timestamps and durations are used. The anonymous data is also used to validate the model and determine different patterns that can be helpful for forecasting demand. Lastly, a method for determining staffing levels is explained for this specific model.

This paper is organized as follows. Section II discussed related work done on either call center models or suicide prevention helplines. Section III describes different patterns found in the data. In Section IV, the proposed model is described. In Section V, the model is validated using a combination of data and trace-driven simulation. Next, Section VI explains how the demand call volumes can be predicted based on historical data. Section VII describes how the model and the forecast can be used to construct staffing advice. Finally, in Section VIII, the conclusion and discussion are given.

II. Related work

Call centers and/or helplines have been researched in different fields of science. For this paper, we will mention some of the research done on the modeling of call centers and research done on specifically suicide prevention helplines. Various research has been done on modeling call centers, Garnett et al. [21] have shown how and why Erlang-B and Erlang-C lack the feature of impatient customers. They introduce an Erlang-B model, where customers abandon the system after not being answered after their impatience has run out. Here, impatience is drawn from an exponential distribution. This research, however, lacks the inclusion of multi-skill. Gans et al. [19] provide an overview of different types of skill routings, where the skill may be based on training, compensation, or time restrictions. Forecasting of arrivals has also been researched, for example, Gijo et al. [22] have shown how (S)ARIMA models can be used in forecasting the call volume. Research done specifically at mental health helplines is often limited to conversation topics and different types of callers. Salmi et al. [14] show the change in conversation topics during COVID-19. While Grigorash et al. [15] and O'Neill et al. [16] have both studied different caller types and how these caller types can be predicted. This paper aims to fill the gap between research done on standard call centers and that done on mental health helplines, specifically by applying modeling and forecasting, that are normally seen in standard call centers, to mental health helplines.

III. Data Analysis

The data for this research is provided by 113 Suicide Prevention [2], referred to as '113'. Its mission is to prevent suicides and break the taboo surrounding suicide. Help seekers struggling with suicidal thoughts can contact 113 24/7 by either chat or telephone. These help seekers are then helped by counselors consisting of volunteers and paid employees. Apart from that, 113 also provides online therapy, self-tests, and self-help courses [25].

Call data is provided over the period 2016-2021 and consists of around 250,000 chats and 175,000 telephone calls. The distribution of the arrivals over the years can be seen in Figures 1 and 2 for phone and chat, respectively. It should be noted that data for 2016 and 2021 were incomplete at the time, explaining the low number of arrivals in 2016 and 2021. However, omitting 2016 and 2021, it can still be seen that there is an increasing trend present in the number of arrivals of phone calls as well as chats. The number of telephone arrivals increased from almost 20,000 in 2017 to more than 50,000 in 2020, and chat arrivals increased from close to 40,000 in 2017 to almost 70,000 in 2020. Each call or chat has the following elements: (1) the arrival time, (2) the time entering the queue, (3) the time of acceptance by an agent, (4) the disconnection time, and (5) the completion time, all elements are denoted using a combination of date and time, using the

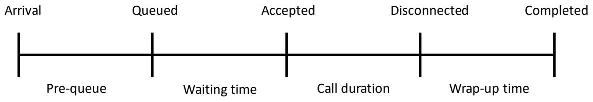

Figure. 3. Timeline of a call from the perspective of the caller.

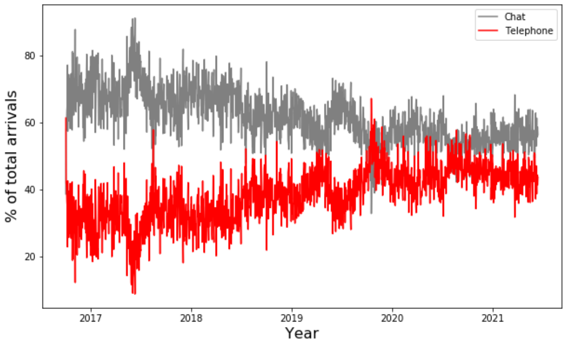

Figure. 4. Decomposition of incoming calls and chats.

following formats dd-mm-yyyy and hh:mm:ss for date and time, respectively. Each call also has a contact id and an *initial* contact id, which may be different if a call is a forwarded telephone call or chat.

Different features were added to the data to obtain more understanding of the different durations found in the data. The following durations were calculated: *pre-queue duration*, *waiting time*, *call duration* and *wrap-up time*. The *pre-queue duration* is the time between the arrival of a call and the time entering the queue, and is the time spent by the caller in a menu. This phase does not require resources from the helpline and, therefore, falls outside the scope of this paper. The *waiting time* is defined as the time between entering the queue and the time that an agent accepts the call. The *call duration* where the agent and the help seeker are actually connected is defined as the time between acceptance of a call and the disconnection time. Finally, after each call, the agent has to fill in a wrap-up form, which is the time between disconnection and completion. A visual representation of this timeline is given in Figure 3.

Recall that there are two options for help seekers to contact the helpline: via *telephone* or *chat*. These two contact types are mostly handled by the same type of agents, but differ in some important aspects. Traditionally, there used to be more chats than telephone calls. However, the difference has diminished in recent years, and the numbers are comparable. This can be seen in Figure 4, which shows that in 2017 and 2018, most of the arrivals were chats. In 2019, the difference between chat and phone diminished, while since 2020, there are still more chats, but the difference is not as large as before. However, the chat and telephone calls do follow different patterns, which are shown in Figures 7 and 8. Figures 5 and 6 show the weekly patterns; in both cases, the weekends see a lower number of calls. However, for chats, we see a clear dip on Saturday and a small increase on Sunday. We also observe that most telephone calls arrive between 12:00 and 20:00, while chats have a clear peak at 20:00. Both telephone and chat call arrivals show a decrease during the night and early morning until around 5:00 in the morning.

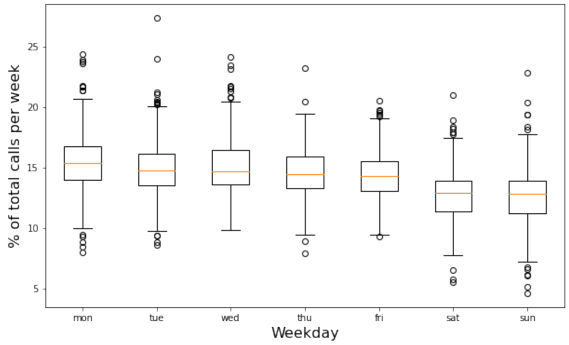

Figure. 5. Weekly pattern of incoming telephone calls.

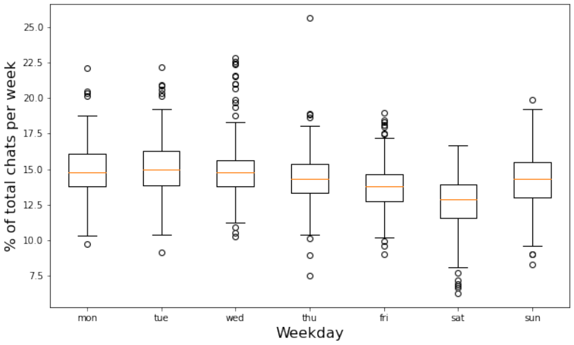

Figure. 6. Weekly pattern of incoming chats.

Incoming chats are first handled by triage, this is needed to filter chats that are at the wrong helpline. Only part of the incoming chats after triage gets sent to another agent. The percentage of chats that an agent handles after triage is around 50% during day shifts, but this differs over the day. During the night shifts, fewer chats are forwarded to an agent. The different nature of night conversations may cause this. The triage plays an important role in filtering chats, as seen by the percentage of chats that get through triage. Chats are filtered out due to various reasons. For example, the chatter may be at the wrong helpline and or is identified as a prank chatter [26].

In the data, we distinguish three service time distributions: for the duration of (1) telephone calls, (2) chats *during* triage, and (3) chats *after* triage. The empirical distribution of phone call durations can be found in Figure 9. On average, a telephone call takes around 18 minutes, with the longest phone call in the data taking multiple hours. The duration of chats in triage can be seen in Figure 10, and the duration of chat conversations after triage can be seen in Figure 11. As can be seen, the chats that have gone through triage tend to take much longer than the chats during triage: on average, chats in triage take 19 minutes versus 34 minutes after triage. This is

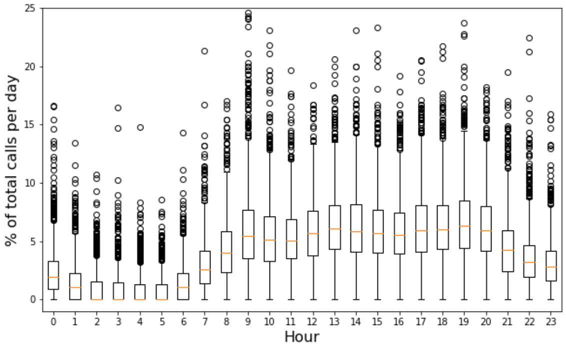

Figure. 7. Daily pattern of incoming telephone calls.

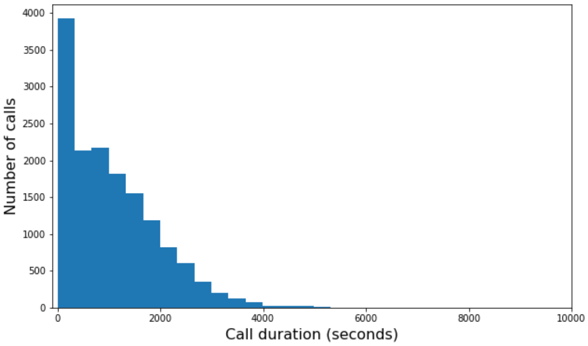

Figure. 9. Histogram of telephone call durations.

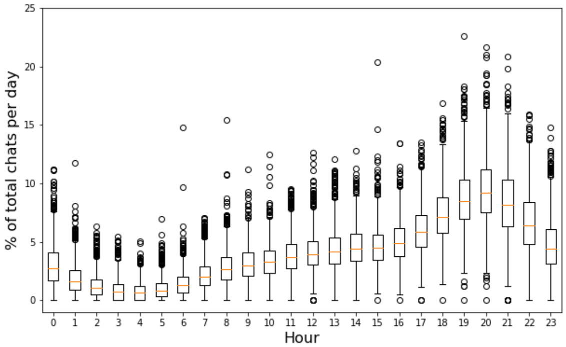

Figure. 8. Daily pattern of incoming chats.

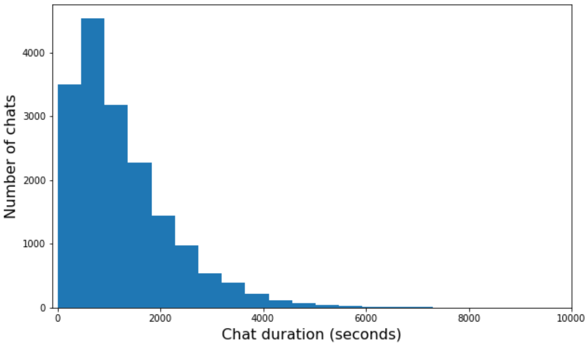

Figure. 10. Histogram of chat triage durations.

also clearly visible in the histograms with durations of chats in triage having a clear peak at around 400 seconds, while chats after triage have a peak in durations at 4000 seconds.

Remark: The effect of the experience of agents
Mental health helplines often use a mix of paid professionals and volunteers, where an agent's responsibilities depend on experience and other factors. The assumption is that experienced agents can handle more difficult conversations alone, while inexperienced agents might require assistance when handling a more complex conversation (this could be either a phone or chat conversation), resulting in a longer service time. Analysis of our data pointed out that there seems to be no significant difference in service time distribution between different volunteers and professionals. There might be a difference, but this could be obscured due to many other factors, such as experienced agents handling the more complex conversations and assisting or coaching the less experienced agents. Based on this observation, the call duration distributions are assumed to be the same for all agent experience levels.

IV. Model description

For the modeling step, we distinguish between two different types of calls, namely: (1) *chat calls*, and (2) *telephone calls*, which are handled differently. The incoming chat calls are first handled by a triage system. The triage system consists of c_{triage} triage agents. Each triage agent can handle n_{triage} chats simultaneously without perceivable slow-down. Together, there are $c_{\text{triage}} * n_{\text{triage}}$ 'triage slots' for incoming chats. Arriving chats enter service at triage immediately if a triage slot is available. If no triage slot is available, then chats enter an infinitely sized queue and are helped first-come-first-served (FCFS) or abandon the queue if the waiting time is longer than their impatience. The service time of a chat call in triage, denoted B_{chat}, is the convolution of four random variables:

$$B_{\text{chat}} = B_{\text{warm-up}} + B_{\text{conversation}} + B_{\text{wrap\_up}} + B_{\text{cool-down}}.$$

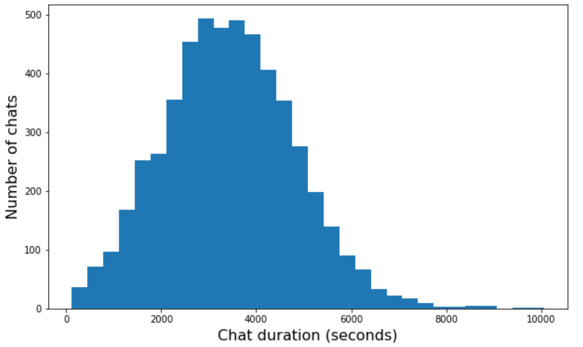

Figure. 11. Histogram of durations of chat conversations after triage.

All variables are drawn from some probability distribution, which can be estimated from the data. A visual representation of the service time can be seen in Figure 12.

A key aspect of the model is the inclusion of *impatience*, i.e., the maximum amount of time that a help seeker is willing to wait before he abandons the system. The impatience of a help seeker who enters the system via a chat is modeled as an independent sample from some probability distribution with mean $\mu_{\text{chat-impatience}}$. After service completion at the triage center, a chat is either sent through to the helpline (HL) for assistance (with probability $p_{\text{sent-through}}$), or the chat leaves the system. Note that during the warm-up period $B_{\text{warm-up}}$, the agent is busy, but the help seeker is not yet answered. Therefore, help seekers may abandon the queue during that period.

Different from chat calls, incoming telephone calls do *not* go through a triage phase and arrive directly at the HL. This is the core part of the system where most of the service processing occurs. The HL is equipped with c_{HL} agents, each of which can handle both chat calls and telephone calls, not more than one call at a time. When a telephone call finds an HL-agent available, he enters service immediately. If the telephone call arrives and no agent is available, the call enters an infinite-sized queue that is handled on an FCFS basis. Here, phone calls are able to abandon the queue if their waiting time is longer than their impatience.

The HL processes both telephone calls and forwarded chat calls (i.e., those that have passed through the triage phase). Here, *chat calls have non-preemptive priority over telephone calls*. Thus, when an HL-agent becomes idle, he first checks whether there is a chat call pending (while keeping a triage slot occupied), and if so, starts to service the longest-waiting chat call. If no chat call is pending, the agent checks if telephone calls are pending.

Similar to the modeling of chat sessions in triage, the duration of phone calls and chats after triage both consists of four subsequent independent phases: (1) warm-up, (2) conversation, (3) wrap-up, and (4) cool-down, where each phase has its own probability distribution differently for phone and chat after triage. The impatience of help seekers via telephone is modeled as a sample from some probability distribution that can be obtained from the data. Calls abandon the queue if their waiting time exceeds the impatience. When service is completed, the calls exit the system. Note that agents are busy during the warm-up period, similar to abandonments at the triage, but the help seeker does not perceive this and can, therefore, still abandon the queue during the warm-up phase. See Figure 13 for an illustration of the model.

V. Model validation

The model described in the previous section has to be validated to confirm our modeling choices. The validation was done using trace-driven simulation. For the simulation of this model, the following values are needed for each arrival:

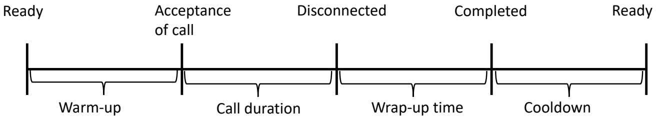

Figure. 12. Timeline of call from the perspective of the agent.

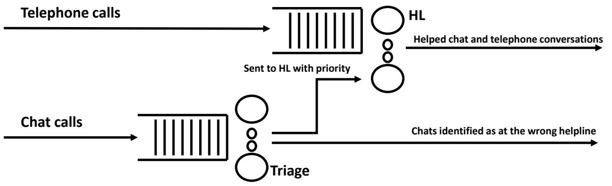
Telephone calls
HL
Helped chat and telephone conversations
Sent to HL with priority
Chat calls
Chats identified as at the wrong helpline
Triage

Figure. 13. Illustration of the model.

(1) an arrival time, (2) service duration (consisting of warm-up, conversation, wrap-up and cooldown), (3) impatience, and (4) call type (chat or telephone) and if the call type is chat information about triage (duration and if the chat is at the right helpline) are also needed. Trace-driven simulation uses values obtained exactly from the data, thereby giving a precise comparison between reality and simulation. The data contained some missing values, and these need to be filled in before simulating, namely: (a) the conversation and wrap-up duration of calls that were unanswered, (b) impatience of help seekers, and (c) warm-up and cool-down durations.

The missing values of conversation and wrap-up durations were filled in using hot-deck-imputation [28]. This method samples from the known values to fill in the missing values. The distinction was made between the different types of conversations: telephone, chats during triage, and chats after triage. The impatience of help seekers are mainly unknown due to the limited availability. Only a small percentage of telephone calls were unanswered, and for chats, even fewer impatience data was available. Warm-up and cool-down durations were not present in the data. Therefore, the missing values of impatience, warm-up, and cool-down periods were all drawn from exponential distributions.

The parameters were estimated using expert opinions from paid professionals and volunteers. Impatience of chat conversation is determined by the sum of a constant 300 seconds and a duration drawn from an exponential distribution with a mean of 300 seconds. The impatience of a telephone caller is drawn from an exponential distribution with a mean of 240 seconds. The warm-up for both chat and telephone is drawn from exponential distributions with means of 60 and 45 seconds for telephone and chat, respectively. Lastly, the cool-down durations of chat and telephone are both drawn from an exponential distribution with a mean of 120 seconds.

Moreover, based on current practice at 113, for the experiments, we assume that $n_{\text{triage}} = 5$. Thus, each triage agent can handle a maximum of five triage chats simultaneously. Further, the simulations are trace-driven, and follow the realizations of the time variations of (1) the arrival processes for chat and

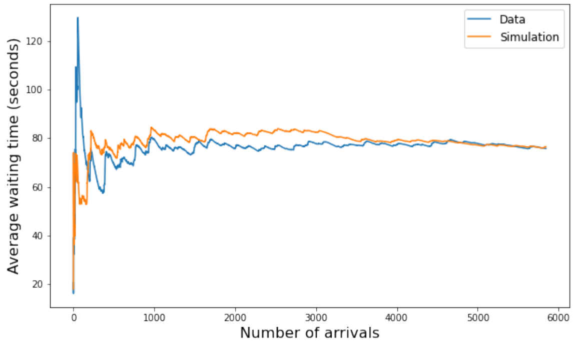

Figure. 14. The average waiting time of telephone calls in the simulation and the data of August 2021.

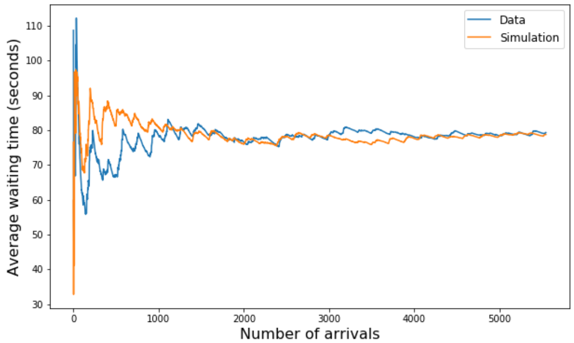

Figure. 16. The average waiting time of telephone calls in the simulation and the data of September 2021.

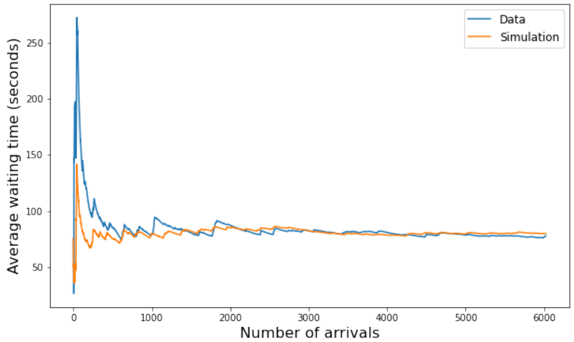

Figure. 15. The average waiting time of chat calls in the simulation and the data of August 2021.

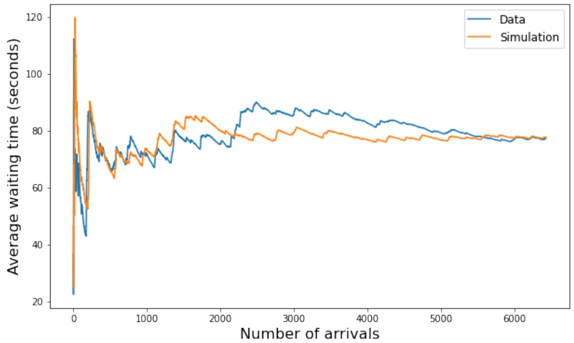

Figure. 17. The average waiting time of chat calls in the simulation and the data of September 2021.

telephone, (2) the number of triage- and HL-agents, and (3) the fraction of triage call that is forwarded to the HL system.

The simulation ran for almost 12,000 arrivals consisting of around 6,500 chat arrivals and 5,500 phone calls during a period of 1 month. The average waiting time of the simulation and the data were then compared. Figure 14 shows both the simulated and realized average waiting times for telephone calls as a function of the cumulative number of call arrivals. The results show that for a small number of calls, the average waiting time is rather sensitive to outliers but quickly stabilizes over time. The average waiting time of the data and simulation both converge to the same value, around 80 seconds. This validation experiment is repeated for chat call arrivals, which show similar convergence, see Figure 15. This process was repeated over different months and showed similar convergences; an example can be seen in Figures 16 and 17, which shows the experiment repeated for September.

In summary, these validation results show that the model works well in predicting waiting times and confirms the modeling choices made in Section IV.

VI. Demand forecasting

Demand forecasting concerns itself with predicting the call volumes for both telephone and chat calls. The performance of the forecasts is dependent on several choices: the time window and the aggregation level. The time window concerns itself with how long ahead the forecast is. Forecasting daily volumes for tomorrow is often easier than the daily volume over 8 weeks. For this paper, we chose to predict 1 day ahead until 8 weeks ahead, since the schedule of the agents is made 8 weeks ahead but can be adapted over time. The aggregation level concerns itself with the kind of volumes to be forecasted. Hourly volumes are harder to predict than monthly volumes, but are less useful for scheduling. In this paper, we chose to predict daily volumes. However, later in this section, it is explained how these daily volumes can be adapted to hourly volume predictions. The choice was made to forecast chat and telephone arrivals independently due to the difference in arrival processes (also explained in Section III) and the difference in handling.

The following forecasting models were considered:(1) Long Short-Term Memory (LSTM), (S)ARIMA [29], linear regression [30], and different baseline models. The parameters of (S)ARIMA models are chosen using auto-ARIMA [31], which are $(5, 1, 1)$ for ARIMA and $(1, 1, 1)(0, 1, [1, 2], 7)$ for SARIMA. As baseline, the prediction of day i is the volume measured on day $i-7$ (baseline model 1) and $i-58$ (baseline model 2). The following aspects were seen as important: *trend* and *seasonality*. The provided data shows that there is an increasing trend present and that weekly cycles seem to be predominant. Therefore, it is chosen to focus on forecasting

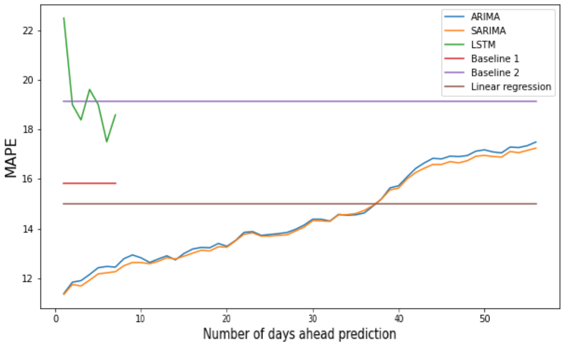

Figure. 18. The MAPE error when forecasting telephone.

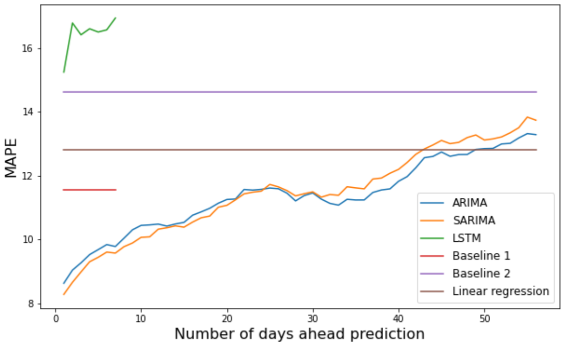

Figure. 19. The MAPE error when forecasting chat.

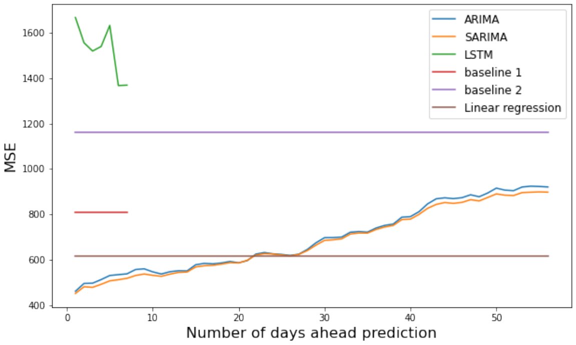

Figure. 20. The MSE error when forecasting telephone.

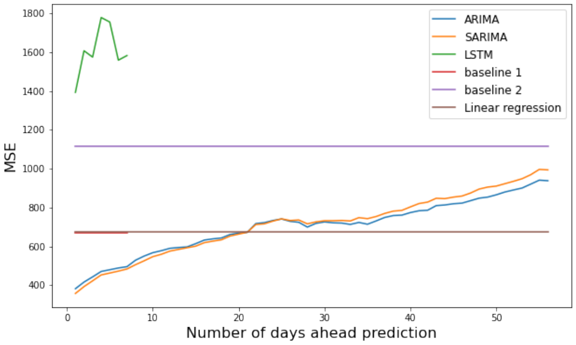

Figure. 21. The MSE error when forecasting chat.

using historical data on call volumes. The models will be evaluated using the Mean Absolute Percentage Error (MAPE), the Mean Squared Error (MSE), and the Mean Absolute Error (MAE). The MAPE is calculated using the following formula:

$$\frac{1}{n}\sum_{i=1}^{n}|\frac{F_i - A_i}{A_i}| * 100\%.$$

Here n is the number of forecasts, F_i is the forecast of day i, and A_i is the actually recorded number of arrivals on day i. The found MAPE results can be seen in Figures 18 and 19. We find that (S)ARIMA models perform best (in terms of the MAPE), especially when forecasting for five weeks (or six in the case of chats) ahead or less. For longer forecasting windows, it turns out that a simple linear regression model might provide more accurate forecasts in the case of telephone arrivals. However, both (S)ARIMA and linear regression models have a MAPE that is lower than the MAPE of the baseline models. The LSTM model has the highest error term in this situation.

Next, the models are evaluated using MSE, which gives more weight to large prediction errors. The MSE is calculated as follows:

$$\frac{1}{n}\sum_{i=1}^{n}(A_i - F_i)^2.$$

The found MSE values can be found in Figures 20 and 21. The results show that in terms of MSE, (S)ARIMA models have the lowest error. However, for forecasting more than three weeks ahead, the linear regression seems to have a lower error term. This differs from the finding when comparing the MAPE error terms, meaning that the (S)ARIMA models likely have more large prediction error than the linear regression.

Lastly, for a complete comparison, the models are also evaluated using MAE, which is calculated by the following formula:

$$\frac{1}{n}\sum_{i=1}^{n}|A_i - F_i|.$$

The found MAE values can be found in Figures 22 and 23. The graphs show similar findings when evaluating based on MAPE, in the telephone calls as well as chat, (S)ARIMA models perform best when forecasting for short-term windows, and linear regression for longer-term forecasts. The MAE that again forecasting chats is more accurate, and the window when (S)ARIMA performs best is longer (around 40 days for chats versus around 30 days for telephone).

Combining the evaluation of the three different error terms, we can come to the following findings:

1) All error terms agree that for short-term forecasting (S)ARIMA models perform best, long-term forecasting

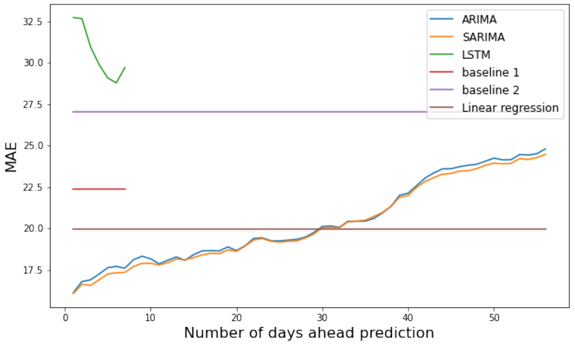

Figure. 22. The MAE error when forecasting telephone.

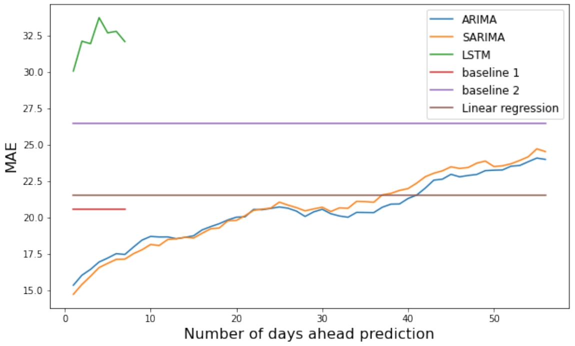

Figure. 23. The MAE error when forecasting chat.

can best be done using linear regression, and LSTM seems to perform the worst.

2) The different error terms differ on the time window of when (S)ARIMA model performs the best. The MSE graph of phone forecasting shows a shorter time window in which (S)ARIMA performs best when compared to that of the MAPE graph. This tells us that it is likely that phone forecasting has more large errors influencing the MSE. A similar picture can be seen for chat forecasting.
3) All evaluations agree that the LSTM model performs the worst of all evaluated models at the moment.

The performance of (S)ARIMA models can be attributed to the flexibility of the models. However, it could be the case that with more time and optimization, the LSTM will perform better. However, it is questionable if the model would perform better than the (S)ARIMA models.

Hourly predictions

The results above concern the prediction of daily volumes. Call centers, however, often use hourly volumes for staffing purposes. Predicting hourly volumes can be done in several ways, namely forecasting hourly volumes directly or indirectly by using the predicted daily volumes. Forecasting hourly volumes directly introduces more uncertainty in the predictions. Therefore, it was chosen to forecast using the daily volumes. The call volume of hour h on day i is calculated as follows:

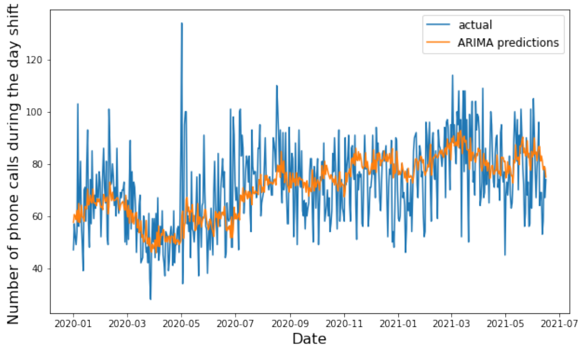

Figure. 24. Comparison between the actual and predicted number of phone calls during the day shift (from 8.00 until 16.00).

$$F_{i,h} = F_i \times P_h.$$

Here P_h is the mean percentage of arrivals on a day that arrives in hour h. This method can be expanded to also include smaller or larger intervals, such as shifts. An example of this can be seen in Figure 24, again the prediction is able to follow the actual number of arrivals.

VII. Staffing

This section will briefly describe how the model of the helpline can be used together with demand forecasting to construct staffing advice. Staffing concerns itself with the required number of agents per time interval and is, therefore, different from a work schedule or planning of agents, for a schedule often staffing advice is first needed. One of the most well-known and often used rules is the square-root staffing rule [32]. This is a formula used to calculate s, the staffing level, and can be rounded up to an integer value. The formula is given by:

$$s = \rho + \beta\sqrt{\rho},$$

where ρ is the offered load calculated by multiplying the arrival rate with the mean service time, β is a parameter reflecting the service level, a higher beta corresponds to a higher quality of service.

The given formula is based on M/M/C queues and can be adapted to other queues. The model described in Section IV needs two different staffing levels at Triage and at HL. The staffing at Triage receives only chat arrivals, ρ_{Triage} is, therefore, calculated by multiplying the chat arrivals with the average service time of chats in Triage. Since staffing is done ahead of time, the chat arrivals are predicted chat arrivals. Together with the fact that agents at Triage can handle five conversations simultaneously, the square-root staffing rule therefore becomes:

$$s_{Triage} = \frac{\rho_{Triage} + \beta \times \sqrt{\rho_{Triage}}}{5}.$$

The staffing at HL can be determined using the square-root staffing rule together with the facts that each agent can handle one conversation and agents handle chats after Triage and new phone calls. The formula then becomes:

$$s = \rho_{HL} + \beta\sqrt{\rho_{HL}}.$$

Here, ρ_{HL} is calculated by summing the chats after Triage multiplied by the average chat service duration, and the phone arrivals with average phone service duration, where again the arrivals are done by forecasting.

VIII. Conclusion

This paper on call center modeling for mental helplines has several contributions. The first contribution is a new call center model for mental helplines. The modeling is based on data and the experience of agents, the model is then validated based on data from [2], the suicide prevention helpline in the Netherlands. The validation is done using trace-driven simulation, and the results show that the model is able to accurately predict waiting-time performance for telephone and chat arrivals at the call center. We emphasize that the model is also applicable to other mental helplines with triage and complex conversations requiring warm-up and cool-down periods, possibly with minor modifications.

The warm-up and cool-down durations are estimated mostly based on experience from agents. There is difficulty in accurately measuring these durations. For future research, these durations could be based on data and possibly correlated to the duration of the call and wrap-up. This could be done by measuring when agents log off and log back on.

In the model, we assume that $n_{triage} = 5$, meaning each agent can handle a maximum of five triage chats simultaneously. Here, we assume that there is no slow down in service time when an agent handles a triage chat more, but is unable to handle a sixth chat. For future research this assumption could be further explored, will the quality decrease and service time increase when setting n_{triage} higher.

The second contribution of this research is centered around demand forecasting. Various models were tested for forecasting the number of arrivals per day and evaluated based on MAPE, MSE, and MAE. All evaluations agreed that for short-term forecasting (S)ARIMA models perform best and linear regression in the case of long-term forecasting. However, the tipping point at which linear regression performs better than (S)ARIMA differs per evaluation. It is also discussed how these daily volume predictions can be changed into hourly or shift volumes.

Lastly, the third contribution is on how to combine the previous two contributions (modeling and forecasting) for staffing purposes. We explain how the model and forecast can be used to adapt the square-root staffing rule, which in turn can be used to generate staffing advice. Regarding the real-life applicability of this paper, the staffing advice and forecasts obtained can easily be combined into a user-friendly dashboard, which can easily be used by the planning department to possibly improve staffing.

It is also important to note that while the staffing advice is important for the performance of the helpline, there are also other ways to reduce the waiting time or increase the percentage of answered telephone and chat calls. Some of these methods are already used by 113 at the moment, such as staff in a different time zone, shortening the menu a new help-seeker has to listen, aiming for shorter conversations and staff extra during the switch moments of different shifts and breaks. For example, 113 Suicide Prevention was able to decrease the waiting time in the night by using call center agents working in a different time zone (Suriname in this case). Another method is shortening the time until a help-seeker enters the queue, by shortening the menu the help-seeker has to listen and fill in, resulting in fewer help-seekers abandoning before entering the queue. Lastly, by aiming for shorter conversations agents are able to handle telephone and chat calls faster, and therefore the overall service time decreases resulting in lower waiting times with the same number of agents or requiring less agents for the same performance.

This paper aims to provide a complete view, but some aspects require follow-up research. For example, the level of experience of volunteers and paid professionals might affect call durations. Preliminary data analysis into this topic shows no, or at best limited, correlation, but further investigation is needed.

Acknowledgment

The authors would like to thank Robin Costers, Sophie de Vries Robles, Menno Zwart, and Kim Setkowski from 113 Suicide Prevention for their great support during this project.

References

[1] T.R. de Boer, S. Mérelle, S. Bhulai and R. D. van der Mei, "A Call Center Model for Online Mental Health Support," in PREDICTIONS SOLUTIONS 2022, International Conference on Prediction Solutions for Technical and Societal Systems, 2022, pp. 1-6.

[2] "About us 113 Suicide Prevention." ("Over ons—113 zelfmoordpreventie.") [Online]. Available: https://www.113.nl/over-113/over-ons (accessed Jun. 24, 2022)

[3] "The listen helpline." ("De luisterlijn.") [Online]. Available: https://www.deluisterlijn.nl/ (accessed Jun. 24, 2022)

[4] "The helpline for children." ("Kindertelefoon.") [Online]. Available: https://www.kindertelefoon.nl/ (accessed Jun. 24, 2022)

[5] J. Scerri, A. Sammut, S. Cilia Vincenti, P. Grech, M. Galea, C. Scerri, D. Calleja Bitar, and S. Dimech Sant, "Reaching out for help: Calls to a mental health helpline prior to and during the covid-19 pandemic," International Journal of Environmental Research and Public Health, vol. 18, no. 9, 2021. [Online]. Available: https://www.mdpi.com/1660-4601/18/9/4505

[6] M. Brülhart, V. Klotzbüucher, R. Lalive, and S. K. Reich, "Mental health concerns during the covid-19 pandemic as revealed by helpline calls," Nature, vol. 600, no. 7887, pp. 121–126, 2021.

[7] J. Hoogenboezem and T. Traag, "Zelfdoding in Nederland: Een overzicht vanaf 1950," Aug 2021. [Online]. Available: https://www.cbs.nl/nl-nl/longread/statistische-trends/2021/zelfdoding-in-nederland-een-overzicht-vanaf-1950?onepage=true

[8] WHO, "Suicide." [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/suicide

[9] M. S. Gould, J. Kalafat, J. L. HarrisMunfakh, and M. Kleinman, “An evaluation of crisis hotline outcomes. part 2: Suicidal callers,” Suicide and Life Threatening Behavior, vol. 37, no. 3, pp. 338–352, 2007.

[10] J. K. Mokkenstorm et al., “Evaluation of the 113online suicide prevention crisis chat service: outcomes, helper behaviors and comparison to telephone hotlines,” Suicide and Life-Threatening Behavior, vol. 47, no. 3, pp. 282–296, 2017.

[11] N. Izady and D. Worthington, “Setting staffing requirements for time dependent queueing networks: The case of accident and emergency departments,” European Journal of Operational Research, vol. 219, no. 3, pp. 531–540, 2012.

[12] M. E. Pratt, “The future of volunteers in crisis hotline work,” Ph.D. dissertation, University of Pittsburgh, 2013.

[13] R. C. W. J. Willems, C. H. C. Drossaert, P. Vuijk, and E. T. Bohlmeijer, “Mental wellbeing in crisis line volunteers: understanding emotional impact of the work, challenges and resources. a qualitative study,” International Journal of Qualitative Studies on Health and Well-being, vol. 16, no. 1, 2021, pMID: 34694979. [Online]. Available: https://doi.org/10.1080/17482631.2021.1986920

[14] S. Salmi, S. Mérelle, R. Gilissen, R. D. van der Mei, and S. Bhulai, “Detecting changes in help seeker conversations on a suicide prevention helpline during the covid- 19 pandemic: in-depth analysis using encoder representations from transformers,” BMC public health, vol. 22, no. 1, pp. 1–10, 2022.

[15] A. Grigorash, S. O’Neill, R. Bond, C. Ramsey, C. Armour, M. D. Mulvenna et al., “Predicting caller type from a mental health and well-being helpline: analysis of call log data,” JMIR Mental Health, vol. 5, no. 2, p. e9946, 2018.

[16] S. O’Neill, R. R. Bond, A. Grigorash, C. Ramsey, C. Armour, and M. D. Mulvenna, “Data analytics of call log data to identify caller behaviour patterns from a mental health and well-being helpline,” Health informatics journal, vol. 25, no. 4, pp. 1722–1738, 2019.

[17] F. Sundram, T. Corattur, C. Dong, and K. Zhong, “Motivations, expectations and experiences in being a mental health helplines volunteer,” International journal of environmental research and public health, vol. 15, no. 10, p. 2123, 2018.

[18] M. D. Christian, “Triage,” Critical care clinics, vol. 35, no. 4, pp. 575–589, 2019.

[19] N. Gans, G. Koole, and A. Mandelbaum, “Telephone call centers: Tutorial, review, and research prospects,” Manufacturing & Service Operations Management, vol. 5, no. 2, pp. 79–141, 2003.

[20] T. R. Robbins, D. J. Medeiros, and T. P. Harrison, “Does the erlang c model fit in real call centers?” in Proceedings of the 2010 Winter Simulation Conference. IEEE, 2010, pp. 2853–2864.

[21] O. Garnett, A. Mandelbaum, and M. Reiman, “Designing a call center with impatient customers,” Manufacturing & Service Operations Management, vol. 4, no. 3, pp. 208–227, 2002.

[22] E. Gijo and N. Balakrishna, “Sarima models for forecasting call volume in emergency services,” International Journal of Business Excellence, vol. 10, no. 4, pp. 545–561, 2016.

[23] M. van Buuren, G. J. Kommer, R. D. van der Mei, and S. Bhulai, “EMS call center models with and without function differentiation: A comparison,” Operations Research for Health Care, vol. 12, pp. 16–28, 2017.

[24] M. van Buuren, G. J. Kommer, R. D. van der Mei, and S. Bhulai, “A simulation model for emergency medical services call centers,” in 2015 winter simulation conference (WSC). IEEE, 2015, pp. 844–855.

[25] M. C. A. van der Burgt, S. Mérelle, A. T. F. Beekman, and R. Gilissen, The impact of COVID-19 on the suicide prevention helpline in the Netherlands. Crisis. 2022.

[26] A. Weatherall, S. Danby, K. Osvaldsson, J. Cromdal, and M. Emmison, “Pranking in children’s helpline calls,” Australian Journal of Linguistics, vol. 36, no. 2, pp. 224–238, 2016.

[27] R. Whitley, D. S. Fink, J. Santaella-Tenorio, and K. M. Keyes, “Suicide mortality in Canada after the death of Robin Williams, in the context of high-fidelity to suicide reporting guidelines in the Canadian media,” The Canadian Journal of Psychiatry, vol. 64, no. 11, pp. 805–812, 2019.

[28] P. Verboon and E. Schulte Nordholt, “Simulation experiments for hot deck imputation,” Statistical Data Editing, Methods and Techniques, vol. 2, pp. 22–29, 1997.

[29] S. Siami-Namini, N. Tavakoli, and A. S. Namin, “A comparison of ARIMA and LSTM in forecasting time series,” in 2018 17th IEEE international conference on machine learning and applications (ICMLA). IEEE, 2018, pp. 1394–1401.

[30] B. M. Pavlyshenko, “Machine-learning models for sales time series forecasting,” Data, vol. 4, no. 1, p. 15, 2019.

[31] R. J. Hyndman and Y. Khandakar, “Automatic time series forecasting: the forecast package for R,” Journal of statistical software, vol. 27, pp. 1–22, 2008.

[32] L. V. Green, P. J. Kolesar, and W. Whitt, “Coping with time-varying demand when setting staffing requirements for a service system,” Production and Operations Management, vol. 16, no. 1, pp. 13–39, 2007.