Vrije Universiteit Amsterdam

Research Paper Business Analytics

Rule-based Traffic Management for Inbound Call Centers

Auteur:
Tim Steinkuhler

Supervisor:
Prof. Dr. Ger Koole

October 7, 2014

Contents

Preface

This paper was written in the form of a Research Paper Business Analytics, which is part of the curriculum for the Master of Science in Business Analytics program at the Vrije Universiteit Amsterdam.

Currently, in practice, most call centers have a Traffic Management department that ensures performance targets are reached each day by responding to deviations from expected traffic. In the described research, we experiment with Rule-Based Traffic Management at inbound call centers. The goal is to explore the possibilities of performing Traffic Management without the intervention of a human being.

I would very much like to thank Prof. Dr. Ger Koole for the support and advice he has given me during the writing of this research paper. Furthermore, I would like to thank Wout Bakker from han!son for an inspiring talk about traffic management at call centers. Given more time, I would have liked to have implemented more of the aspects we discussed, that would make the results of this research more directly applicable in practice. Moreover, I would like to especially thank Rogier Maarse from SNS Klantenservice for supplying the data used in this research and sharing his call center experience.

Abstract

Call center traffic is subject to uncertainty, causing the performance of call centers to fluctuate. To counter these fluctuations, call centers often have Traffic Management department with people that try to ensure performance targets are reached, by the end of each day.

This research paper describes an experimental approach to implement computer-controlled traffic management at an inbound call center in the form of Rule-Based Traffic Management. A Discrete-event simulation set-up is explained, after which simulations are run to test the effect of three sets of rules against each other and a non-traffic management scenario.

Implementing Rule-Based Traffic Management under our settings, is shown to have a positive effect on call center performance, but against higher costs. For practical implementation, our model should be altered to better fit the situation of the call center at hand.

1 Introduction

Currently, there has been a great deal of research about call centers and the way they should operate in order to achieve optimal results, some of which is described in section 2. In practice, it is impossible to exactly forecast the amount of traffic (requests for service) that arrives at a call center beforehand. Depending on the accuracy of the forecasts, call center performance will deviate from what is required by higher management. To counter the effects of deviations from forecasts, it is customary for call centers to have a traffic management team that makes sure the targets are reached by the end of the day. This study aims to investigate the possibility of replacing traffic management teams by a simple, but objective and consistent, rule-based traffic management system.

Doing so, we aim to answer the following research questions:

- How could Rule-Based Traffic Management be implemented?
    1. How can one determine the right initial staffing levels?
    2. When should Traffic Management take action?
    3. Under what rule(s), should Traffic Management take action?

In the next section, we will describe a selection of previous studies about call centers, the way their personnel is managed, and how Traffic Management comes into play. In section 3, we will explain our discrete event simulation that was used to study the effect Rule-Based Traffic Management. Consequently, in section 4, the results of the simulations are described. In section 5, we will conclude our research, discuss it's limitations and opportunities for further research.

2 Literature Research

Call centers and their contemporary successors, contact centers, have become a preferred and prevalent means for companies in different industries worldwide, to communicate with their customers. These call centers provide a primary link between customer and service provider (Gans et al, 2003). Typically, 60-70 percent of the total operating costs for a call center are the workforce costs (Gans et al, 2003). This has caused Workforce Management (WFM) to be a well represented subject of studies.

2.1 Workforce Management

Call center WFM is about accurately translating demand for service into demand for workforce and finding the optimal service level to personnel trade-off (Koole, 2013). Chen (2014) has translated the process of matching workforce staff to demand for services into four general steps:

1. Choose a service level objective for inbound call centers, e.g. 80% of calls answered in 20 seconds.
2. Forecast the call load in each time block. The forecast includes forecasting the cal arrival rates and the estimation of the call handling time distribution.
3. Calculate staffing levels, i.e., for each time block, calculate the number of staff needed to meet the service level objective according to the forecast call load.
4. Schedule staffing shifts based on staffing levels with the rostered staff factor, shrink factor or shrinkage. This takes into account breaks, training and non-phone work.

2.2 Service Level Objective

The chosen service level objective will differ between different industries and among different companies within the same industry. The call center industry standard for inbound call is to require 80% of all calls to be answered within 20 seconds. Other possible objectives could be to have the average waiting time per call to be below a certain threshold. When setting the service level objective, managers need to make a trade-off between quality and costs (Koole, 2013). Later in this paper, you will find the service level objective that was used for our research.

2.3 Forecasting

Forecasting at call centers is performed based on historical data, to estimate how much traffic (incoming demand for service) can be expected for each

time-period (typically of length 15 or 30 minutes) in a day. In practice, it is found that the arrival processes of calls at a call center are well described by inhomogeneous Poisson processes (Koole, 2013). This means that the time between two incoming calls can be described by an Exponential distribution with rate λ_t, which can differ, depending on the time of the day. Besides variability due to Poisson fluctuations, the number of incoming calls is influenced by Seasonality, Holidays, Day of the week, Actions & Special events and trend (Koole, 2013). Forecasts are often made for different timeframes (e.g. T+1Y for long-term planning, T+5W for monthly scheduling and T+1W for adjusting the weekly schedule), depending on preference per company.

2.4 Workforce Scheduling

Based on calculated forecasts, call center management needs to make sure there are enough agents (call center operatives that process the calls) present during each time period, to meet the chosen service level objective. Analytically, the number of agents required to meet the service level objective can be calculated using the Erlang C formula. This formula calculates the average service level over infinitely many calls, and makes the following assumptions (Koole, 2013):

- Poisson arrivals;
- Exponential service duration;
- A fixed number of undistinguishable agents;
- All calls wait in queue until they get served (infinite patience);
- Calls are answered in order of arrival (longest-waiting call first)

In practice, due to variability, the service level will always vary from the service level calculated by the Erlang C formula. Roubos et al (2011b) provided a method to incorporate service level variability in the model, by showing that the service level is normally distributed around the mean (Erlang C value) and calculating the variance. Using that information, allows one to calculate staffing levels, such that with probability X, at least proportion Y of all incoming calls get served before Z time units pass (when using the standard Erlang formula, $X = 0.5$). The Erlang C formula has been known to consistently overstaff, that is, the actual service level is often higher than the Erlang prediction. The main cause of this, is that no person calling a call center will prove to have infinite patience, therefore a proportion of the calls that are not answered directly, will abandon. To account for abandonments, one can use the Erlang A model described by Koole (2013).

2.5 Shift Scheduling

After determining the required amount of agents for each time-period of the day, working schedules are made. Since no agent will be willing to come to work for just one time-period, agents working at a call center work in shifts. These typically last for eight hours, plus breaks in between. These shifts almost never allow for the number of scheduled agents to perfectly match the required number of agents, which causes overstaffing in certain periods. This type of overstaffing is call *shift inefficiency* (Koole, 2013). Besides from shift inefficiency, it also happens that agents that are scheduled to work, are not available to answer calls or do other work, perhaps because of illness, vacation or because they're receiving training. When scheduled agents are not available to answer a call, this is called *shrinkage*.

2.6 Traffic Management

Call center workforce management does not stop after setting the desired service level, making forecasts of incoming calls, determining the right amount of staffing for each of the studied time-periods and scheduling shifts to match the staffing plan. In everyday situations, the actual number of incoming calls will almost never exactly match the number of calls that were forecasted (either due to bad forecasting or expected Poisson variability). Fluctuations in the actual number of calls will cause the service level, along with other performance measures to fluctuate as well. And even if the forecast is spot on, unforseen (lack of) shrinkage can form another source of fluctuations in performance. Traffic Management (or Real-Time Performance Management, or RTPM) can be seen as the adaptations made to the plans to achieve the right SL and efficiency objectives by the end of the day (Koole, 2013).

Thus, traffic management involves monitoring the SL and other efficiency objectives. When one or more of the objectives are not met, this is either due to understaffing or overstaffing. When one of the two occurs, the right actions need to be determined and taken. Correcting for a low SL due to understaffing, should be done for at least these three reasons (Koole, 2013):

1. To compensate for the low SL in the beginning of the day.
2. To account for the redials that will occur as a consequence of the bad SL.
3. To account for an expected increase in incoming calls. If the reason for the bad SL is that more traffic arrived than was expected, this might well also be the case throughout the rest of the day.

Furthermore, in case of overstaffing, the call center will incur unnecessary high costs for employing idle agents. In this case, the exact opposites of the three rules described above will apply.

Traffic management can be done considering different horizons. For instance, as more information becomes known, traffic forecasts are adapted to this information (e.g., a five-week forecast is translated to a 1-week forecast, incorporating the latest information). In this research, we investigate ways to implement intraday Traffic Management to avoid under- and overstaffing, consequently maintaining the required quality of service, whilst keeping down costs. In the following section, we will explain our research model, the methods used to evaluate model performance and consequently, we will descibe the outcomes of this research.

3 Research Methods

In this section, we will first describe the settings of the discrete event simulation that is used in our research. Consequently, we will explain the different scenario's under which our traffic management rules are tested, followed by the characteristics of the traffic management rules and the evaluation methods used to establish success.

3.1 Discrete Event Simulation Model

To estimate the effect of our traffic management model, we chose to use discrete event simulation. The results of the simulation are designed to closely resemble the actual outcomes of a 13-hour long workday at a call center that receives inbound calls. To determine the parameters for our model, we used data from a Dutch financial institution. This data contained the forecasts and actual traffic for Thursday, September 18th 2014 and estimates of the average patience and handling time, along with the service level objective. In this research, only inbound calls are modelled. We will now describe the different aspects that describe the simulation settings.

3.1.1 Call Arrival Process

Call arrivals are generated by a pseudo-random number generator. The arrivals are assumed to follow an inhomogeneous Poisson process. Consequently, inter-arrival times are given by an exponential distribution with rate λ_t, associated with time-period t. These rates for λ_t are initially assumed to be well-described by the set of intraday forecasts that is described in figure 1. The figure shows the proportion of the total (3124) daily calls per 15-minute time-period.

In this scenario, the only fluctuation in the results of the simulations, will be due to Poisson variability. To incorporate the effect of deviations forecasting errors, a busynessfactor is introduced as proposed by Whitt (1999). In each independent run (or simulated day), a random busynessfactor B is drawn from a distribution with $E(B) = 1$. B then represents the busynessfactor for that day and is used to calculate new rates for the inhomogeneous Poisson process that generates incoming calls. As distribution for B, we chose to take a Normal distribution with mean 1 and standard deviation 0.1. This means on average, more than 3 out of 10 days, the forecast will be off by more than 10%. The distribution of B should be altered to fit the required level of deviations from the forecasts. The rates are then calculated as:

$$B \sim Norm(1, 0.1^2)$$

$$\Lambda_t = B * \lambda_t, \forall t \in \{0, \ldots, T\}$$

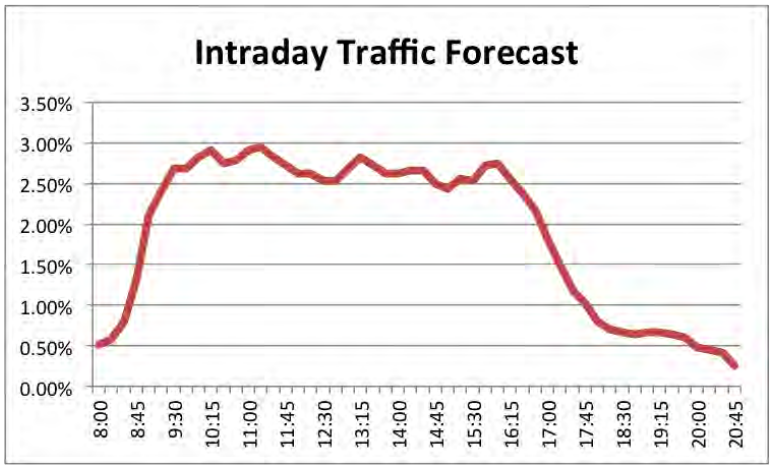


Figure 1: Forecasted percentage of daily traffic per 15 minute interval

B is generated once for each simulated day, meaning all per-time-period arrival rates in one simulation are multiplied by the same busynessfactor.

3.1.2 Abandonments

In any real world situation, people calling a call center will not have time to wait in a queue for an infinite amount of time. That is, no one will have infinite patience and therefore at some point in time, when people are left waiting, they will hang up (or abandon). However, not every person will have the exact same length of patience. In the simulation, each call that is not directly answered, is attributed a patience of length p, drawn from an Exponential distribution with rate 0.4 ($E(p) = 2.5$). For patience calculations, see Koole, 2013. Once the waiting time for a call exceeds p, the call is removed from the queue and labeled as an abandoned call.

3.1.3 Call Length Distribution

Each incoming call requires a certain amount of time from an agent to process. This amount of time is called the handling time and includes talking time and wrap-up time. In our simulation, the average handling time (AHT) is set at 10 minutes and drawn from an exponential distribution with rate $\mu = 0.1$. We note that in practice, the handling time per call may vary per agent and per time of the day. We did not model this. We assume that all agents are equally fast and are expected to treat calls in the same way, regardless of what time it is.

3.1.4 Perfomance measures

The goal of the call center in our research is to achieve a 80% SL, with an acceptable waiting time (AWT) of 60 seconds and to do so, for the lowest cost possible. Hence, a Traffic Management system is said to perform better, when the resulting Service Level (SL) is higher, the target service level (TSL) is reached with greater probability, and / or if the costs per connected call are lower.

The SL for a simulation run is calculated as SL_2 from Koole (2013, page 23):

$$SL = \frac{\#(connected \leq AWT)}{\#(connected) + \#(abandoned > AWT)}$$

Agents are all assumed to cost 1 unit per hour and in case of deviations from the planned schedule, extra costs are incurred. Adding an agent to the schedule will cost 1.1 units per hour and removing an agent, will cost 0.1 units per hour. In practical situations, these costs may be adjusted to better reflect reality.

3.1.5 Other Simulation Assumptions

Furthermore, the simulation assumes there is unrestricted capacity to scale the amount of agents scheduled to inbound calls, by assigning agents extra to inbound calls (scaling up) or assigning agents to other tasks (scaling down). Hence, any shrinkage will be compensated for and is not modelled. In a less experimental setting, capacity and shrinkage should be introduced to the model.

3.2 Initial Staffing

Traffic Management can be seen as the adaptations made to the plans to achieve the right SL and efficiency objectives by the end of the day (Koole, 2013). Hence, before testing the performance of different traffic management systems, there need to be initial plans. To determine the initial staffing levels per time-period, estimated to be required to handle incoming traffic, we compare two models: the traditional Erlang C model and the Erlang A model described by Koole (2013).

To determine the level of staffing needed, the Erlang C and Erlang A formulas were used to calculate the number of agents required to have the expected service at minimum be the target service level (in this case 80% of the calls are required to be answered within 60 seconds). Both the Erlang C and A formulas are based on mathematical models of the call center. The Erlang C formula uses forecasts for the calls, the average handling time and the number of agents to calculate the expected Service Level. The Erlang A formula also takes into account a patience distribution.

Both the Erlang C and Erlang A models are used to calculate steady state probabilities, hence generate average values over an infinite number of calls. In a simulation and in practice, an infinite number of calls will never be reached, consequently simulated and actual results will deviate from the expectation.

In section 4.1, as an answer to our first research question, we describe the performance of both the Erlang C and A models, based on 1000 simulated days at the call center, using both the Erlang C and Erlang A staffing levels. More advanced methods to determine initial staffing levels can be introduced to improve performance, e.g., in Ding (2014), an analytical approach is described to determine cost-optimal staffing in the presence of traffic management costs.

3.3 Traffic Management Timing

Our second research question is: *When should Traffic Management take action?* in section 4.2, we will describe our search for the best moment of the day to perform (or to start performing) traffic management actions. The goal is to try and determine what is the right moment to (start) alter(ing) the original staffing levels to match deviations of actual traffic from the forecasts. Several simulations will be run and the intraday performance will be studied.

3.4 Traffic Management Rules

Our third and final research question is: *Under what rule(s), should Traffic Management take action?* in this research, there are two ways traffic management can take action:

Scaling Up Adding one extra agent to the schedule for each remaining time-period until the end of the day

Scaling Down Removing one agent to the schedule for each remaining time-period until the end of the day

For practical implementation, more actions can be thought of, depending on the setting, to incorporate restrictions on the minimum or maximum amount of time one agent should spend on a certain task consecutively.

We will now describe the rules that are used to determine if one of these actions should be taken. In theory, one can think of many different rules (or combinations of rules) and implement these with many different parameters. The following rules are meant to illustrate and test the concept of rule-based traffic management.

3.4.1 TM 1: Bounded Service Levels

Scale up when: $SL < 80\%$

Scale down when: $SL > 85\%$

In practice, management is punished when the TSL is not reached, hence we choose to scale up when the SL is below the TSL (80%). The upper bound for the SL is chosen to be 5% above the TSL.

3.4.2 TM 2: Restricted Waiting Times

Scale up when: Expected Remaining Waiting Time for the call at the end of the queue is greater than one minute (AWT)

Scale down when: Expected Remaining Idle Time for agent that last finished a call is greater than 2.5 minutes (0.25 * AHT)

Expected Remaining waiting time (W) for the call at the end of the queue, is in this case Hypoexponentially distributed. This means that the remaining waiting time is given by a sum of exponentially distributed variables, each with their own rate λ_i. If k is the total number of calls in the queue, s is the number of agents working, $\mu = 1/AHT$, and $\gamma = 1/p$:

Where we assume s is constant, since it will not happen often that the waiting time is longer than the length of one period (or that the caller is still waiting after that time), thus will not span more than two, and staffing levels will not differ drastically between two consecutive time-periods. $E[W]$ is then calculated as follows:

$$\lambda_i = (s\mu + i\gamma), i \in 0, .., k-1$$

$$E[W] = \sum_{i=0}^{k-1} \frac{1}{\lambda_i}$$

The expected remaining idle time (I) is estimated by using the Erlang distribution with rate λ given by the forecast and shape n given by the total number of idle agents. Where we assume λ and n are constant, for the reason given above. $E[I]$ is then calculated as follows:

$$E[I] = \frac{n}{\lambda}$$

3.4.3 TM 3: Combined Rules

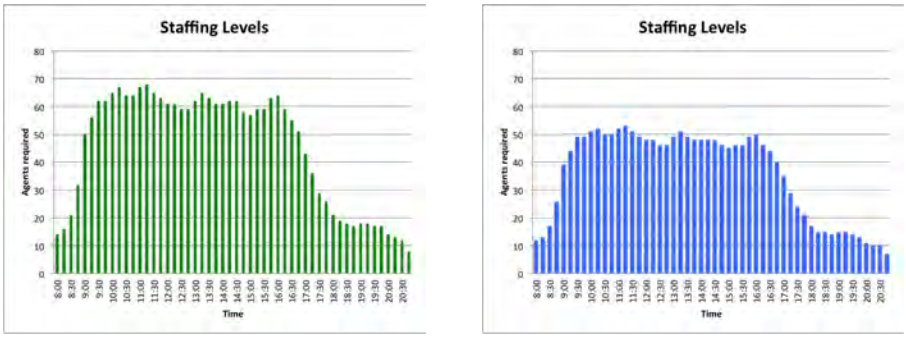
Scale up when: $SL < 80\% \wedge E[W] > 1$

Scale down when: $SL > 85\% \wedge E[I] > 2.5$

Where we combined the rules TM 1 and TM 2, making TM 3 the most restrictive of the three, hence the least amount of actions will be taken using this rule.

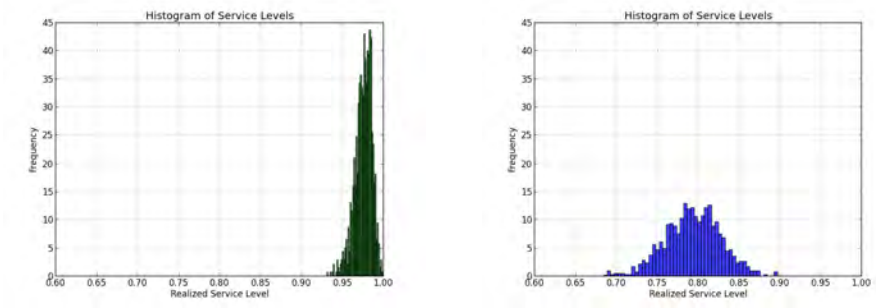
3.4.4 Evaluation

First, we will run simulations (10,000 days each) without implementing any traffic management, with and without using the busynessfactor. For both of these scenario's (we will call these our base-scenarios) we will record the service level and the costs. This will help us understand the possible results of a day at the call center without traffic management interference. Consequently, simulations will be run, with traffic management rules implemented. The results of these simulations will then be compared to the results of the base-scenarios.



(a) Erlang C (b) Erlang A

Figure 2: Required number of agents per time-period



(a) Erlang C (b) Erlang A

Figure 3: Realized Service Levels in case of staffing based on Erlang C or A

4 Results

4.1 Initial Staffing

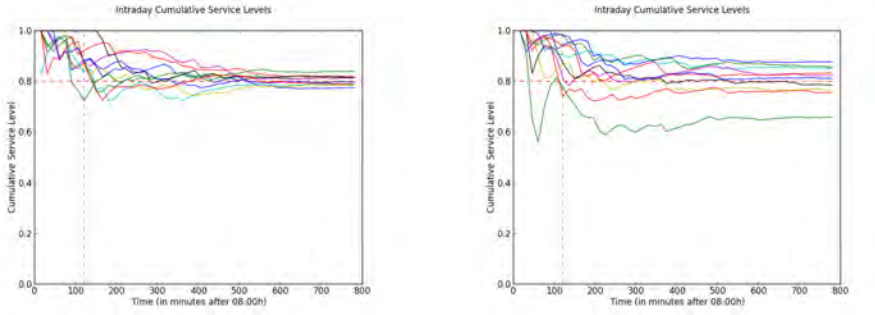
The first research question is: *How can one determine the right initial staffing levels at an inbound call center?* In figure 2, you will find the staffing levels calculated by both the Erlang C and the Erlang A models, based on the forecast that was given in section 3.1. Using these staffing levels, we've run 1000 simulation runs and recorded the service levels. These can be seen in the histograms in figure 3.

As stated in section 2.4, the Erlang C formula does not take patience into account. When simulating 1000 runs with patience as described in section 3.1.2 with scheduling based on infinite patience, the service level of the realizations will be higher than required, as is shown in figure 3a. In fact, in 1000 runs, it doesn't occur that the service level is lower than 92.5%. The average service level that was found with Erlang C staffing is 97.56% ($E[SL_{ErlangC}] = .9756, Var[SL_{ErlangC}] = .00011$).

The Erlang A does take customers' finite patience into account, which

means less agents are required to achieve an the same expected service level (as can be seen in figure 2b). As a result, service levels for the simulated days are also lower. In figure 3b, you can see that the service levels are concentrated around the mean of 79.37% ($E[SL_{ErlangA}] = .7937, Var[SL_{ErlangA}] = .00114$) and approximately follow a Normal distribution.

Using 1000 simulations, we can say with 95% certainty that the actual expected service level will be within a $(79.16\%, 79.58\%)$ confidence interval. For the rest of the simulations, we will assume the initial staffing levels that are given by the Erlang A model (figure 2b).



(a) Busynessfactor $B = 1$ (b) Busynessfactor $B \sim Norm(1, 0.1^2)$

Figure 4: Intraday cumulative service levels for 10 simulations, with and without random busynessfactor

4.2 Traffic Management Moment of Action

The second research question is *When should Traffic Management take action?* To answer this question, we investigated the intraday development of the service level. After setting the initial staffing levels with the Erlang A formula, each simulation starts at 08:00h in the morning with an empty queue, and calls start to come in. An emtpy queue means that agents are available immediately to answer the first few calls and the service level will be 100% when calculated over these first calls. Consequently, traffic management should not act in response to the service level too early on the day. Moreover, traffic management should not act too late either, or there will not be enough time to try and make up for a bad start.

In figure 4, the intraday cumulative service levels are shown, for two settings. On the left, forecasts are assumed to be very accurate (Busynessfactor $B = 1$) and on the right, the busynessfactor is introduced as described in section 3.1.1. It can be seen, that the service level is very volatile at the beginning of the day (single calls have a larger impact on the service level). From about half-way through the day, the service level is more stable. Towards the end of the day, no big shifts can be detected in the service level, this can be explained since there is less traffic near the end of the day (see figure 1) and single calls have less impact on the service level as the total number of calls in the day grows larger.

It would be preferable if the moment of the day could be approximated, at which performance for the rest of the day would become accurately predictable. The results of the simulations lead us to choose for not one, but multiple moments on the day at which the traffic management rules should be evaluated. For the rest of the simulations, the traffic management rule evaluations will take place every 15 minutes after 10:00 h (120 minutes after opening).

4.3 Traffic Management Rules

The third and final research question is: *Under what rule(s), should Traffic Management take action?* In search of an answer to this question, we have simulated 10.000 days at the call center, each time using a different set of the rules described in section 3.4. We will now describe the results of these simulations, that are shown in figure 5 and tables 1 and 2.

4.3.1 No TM

First, when looking at the histograms in figures 5a and 5e, a large spread between realized SLs can be detected, especially in the scenario where B is randomized. Also, based on the top rows in tables 1 and 2, we can conclude that on average, the TSL is not reached, only in 43.5% and 48.31% for the constant and variable busyness scenarios respectively. The total costs without TM are always the same, as no adaptations are made to the work schedule throughout the day. The costs per connected call do vary as a result of random call arrivals.

4.3.2 TM 1: Bounded Service Levels

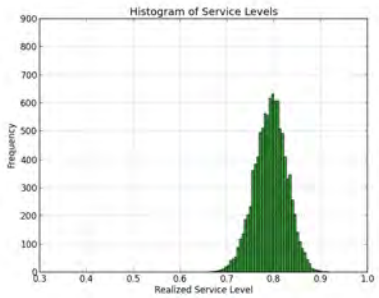
Secondly, implementing TM 1: Bounded Service Levels, causes the realized values for the SL to be both higher, and more concentrated around the mean (see figures 5b and 5f). The TSL is reached in 89.58% and 81.43% for the constant and variable busyness scenarios respectively. Both the resulting average total costs and the costs per connected call are higher than in the scenario without traffic management. This seems to be the result scaling up and down frequently.

4.3.3 TM 2: Restricted Waiting Times

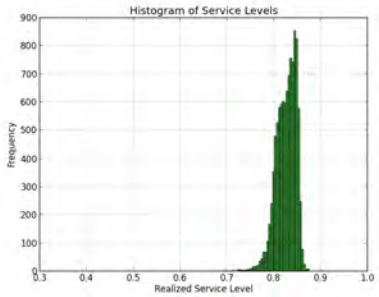
Thirdly, implementing TM 2: Restricted Waiting times causes a higher increase in mean SL than implementing TM 1. However, using TM 2, there is more variance than using TM 1. Furthermore, we can see that the frequency at which the TSL is reached, differs from TM 1; for $B = 1$, it goes up to 95.12%, for $B \sim N(1, 0.01)$, it goes down to 75.48%, when comparing to TM 1. The total costs and costs per connected call are higher than when no TM is implemented, but lower than when implementing TM 1, which is due to scaling up and down less frequently.

4.3.4 TM 3: Combined Rules

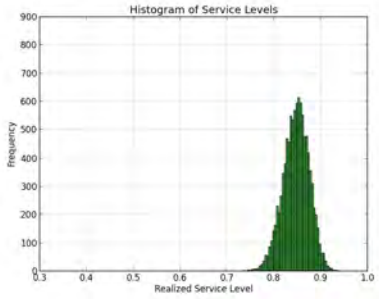
Finally, implementing TM 3: Combined Rules, shows the smallest raise in mean SL out of the three TM rules. But it does so, at the lowest costs and using the least scaling actions.



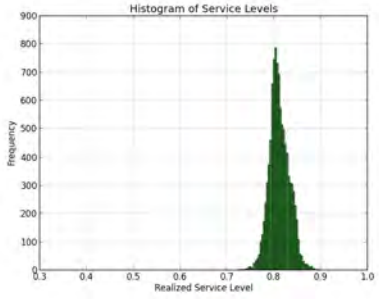
(a) $B = 1$, no TM



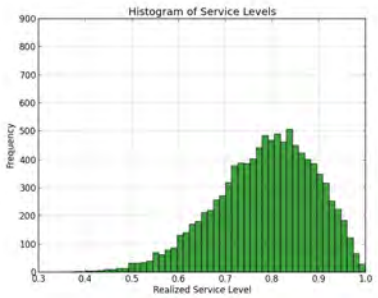
(b) $B = 1$, TM 1



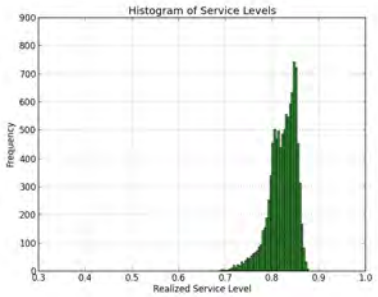
(c) $B = 1$, TM 2



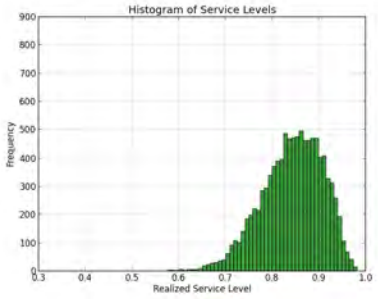
(d) $B = 1$, TM 3



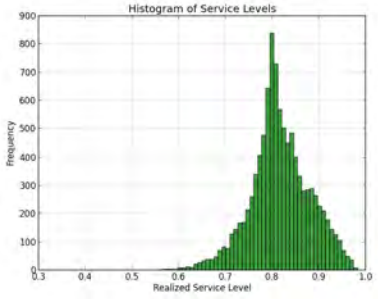
(e) $B \sim Norm(1, 0.01)$, no TM



(f) $B \sim Norm(1, 0.01)$, TM 1



(g) $B \sim Norm(1, 0.01)$, TM 2



(h) $B \sim Norm(1, 0.01)$, TM 3

Figure 5: Histograms of realized end-of-the-day service levels.

$B = 1$	Service Level Mean (Var), $P(SL) > 80\%$	Total Costs Mean (Var)	Costs / Conn. Call Mean (Var)	TM Actions Mean U, Mean D
NO TM	79.35% (0.0011), 43.5%	465.75 (0)	0.1762 (8e-06)	0, 0
TM 1	82.64% (0.0004), 89.58%	512.26 (786.58)	0.1889 (7e-05)	12.69, 7.05
TM 2	84.78% (0.0007), 95.12%	494.27 (109.60)	0.1805 (1e-05)	5.39, 4.90
TM 3	81.18% (0.0004) 70.92 %	476.46 (163.66)	0.1780 (2e-05)	1.72, 0.14

Table 1: Result table for simulations with $B = 1$

$B \sim N(1, 0.01)$	Service Level Mean (Var), $P(SL) > 80\%$	Total Costs Mean (Var)	Costs / Conn. Call Mean (Var)	TM Actions Mean U, Mean D
NO TM	79.35% (0.0011), 43.5%	465.75 (0)	0.1762 (8e-06)	0, 0
TM 1	82.64% (0.0004), 89.58%	512.26 (786.58)	0.1889 (7e-05)	12.69, 7.05
TM 2	84.78% (0.0007), 95.12%	494.27 (109.60)	0.1805 (1e-05)	5.39, 4.90
TM 3	81.18% (0.0004) 70.92 %	476.46 (163.66)	0.1780 (2e-05)	1.72, 0.14

Table 2: Result table for simulations with $B \sim N(1, 0.01)$

5 Conclusion & Discussion

During this experimental research, we have come to somewhat conditional conclusions. First of all, in our simulation setting, the Erlang A model outperforms the Erlang C model, when estimating the optimal staffing levels and requiring the SL to be as close to the TSL as possible.

Furthermore, no strict rules have been found to determine the optimal moment of the day at which one should (start to) implement traffic management. We have not found a closed format to estimate this moment and conclude that it is good to perform traffic management at multiple times during the day and that the moment at which to start with the first traffic management action is highly dependent on the situation. In a practical situation, this moment could be determined based on simulation results, while keeping in mind the variability of the factors affecting the SL.

Moreover, implementing Rule-based Traffic Management can cause an increase in the expected SL, less variance and a greater probability to reach the TSL. However, this often happens at a higher cost. In practical situations, different (combinations of) rules should be tested to estimate the difference between their outcomes, after which a choice should be made considering both the effect on performance and costs. This choice should also depend on the precision of the forecasts.

There are some limitations to our research. In this study, not all aspects of a call center have been taken into account. For our research method to be used in practice, the model should be altered, taking more aspects into account. For example, in practice there will be deviations from 100% adherence. There will also be restrictions on the number of agents that are available for scaling up and down, and the agents will not always be immediately available when called upon. Furthermore, call center agents will not all take the same amount of time for a call, and sometimes they will be unavailable. Moreover, in a typical call center (or contact center), there are more types of activities to be done than handling inbound calls, e.g., handling e-mails, outbound calls or learning skills. The effects of (some of) these additional factor(s) might be studied in further research. Further research might also be done to test other rules and other actions or the effect of other parameters to the rules proposed in this research.

References

[1] Chen, Xi. *Combining Forecasting and queueing models for call centre staffing*, PhD thesis, Lancaster University 2014

[2] Ding, S., Koole, G. *Optimal call center forecasting and staffing under arrival rate uncertainty*, working paper. 2014

[3] Gans, N., Koole, G., Mandelbaum, A. *Telephone Call Centers: Tutorial, Review, and Research Prospects*, Manufacturing & Service Operations Management, Vol. 5, No. 2, pp. 79–141 2003

[4] Koole, G. *Call Center Optimization*, MG Books Amsterdam 2013

[5] Roubos, A., Bhulai, S. & Koole, G. *Flexible Staffing for Call Centers with Non-Stationary Arrival Rates*, working paper. 2011a

[6] Roubos, A., Koole, G. & Stolletz, R. *Service Level Variability of Inbound Call Centers*,Manufacturing & Service Operations Management 14(3), pp. 402–413. 2011b

[7] Whitt, W. *Dynamic Staffing in a Telephone Call Center Aiming to Immediately Answer All Calls*, Operations Research Letters 24: pp. 205–212. 1999