UNIVERSITY OF CALIFORNIA
Los Angeles

Forecasting and Improving the Call Center Operations
Time series approach and
Queueing theory approach

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Applied Statistics

by

Jixuan Li

2018

ABSTRACT OF THE THESIS

Forecasting and Improving the Call Center Operations

Time series approach and

Queueing theory approach

by

Jixuan Li

Master of Applied Statistics

University of California, Los Angeles, 2018

Professor Qing Zhou, Chair

This paper mainly aims to provide the data story to the call center to improve operations, assisting the Business Experience Team to make data-driven decisions. We intend to capture the pattern of current trends and seasonality, as well as the forecasting for the next peak-season call volume by time series approach. More importantly, we would like to identify the optimum staffing levels needed to meet certain service goals by implementing queueing model. In order to predict the baseline and the peak season of the call volume, we compare 2 time series models: ARIMA (autoregressive integrated moving averages) model and Holt-Winters exponential smoothing model. The later one gives a better prediction of the call volume. Moreover, we employ 3 classical models (ErlangC, ErlangB and ErlangA models) to find the relationship between the number of operators needed and the wait time of customers in the queue. We end up with the conclusion that increasing 5 operators could achieve the target: 95% of phone calls are answered within 5 minutes.

The thesis of Jixuan Li is approved.

Yingnian Wu

Nicolas Christou

Qing Zhou, Committee Chair

University of California, Los Angeles

2018

To my mother, Xiao Jun Liu ...

who always support me, whatever path I took

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to thank my supervisor, Nicolas Christou, for his invaluable guidance and helpful advice throughout my thesis. I would like to thank my industrial partner, City of Los Angeles, Office of Finance, providing me with the dataset and supports.

CHAPTER 1

Introduction

The call center provides services via telephone-based services. In the City of Los Angeles, Office of Finance is pummelled with calls in persons inbound, emails and web traffic during the January and February, the city business tax renewal season.

The main challenge is that they need to increase the number of operators to improve the call center traffic. During the peak season, the staff is interested in determining the number of operators to improve the operations at a more advanced level. We can provide data-driven solutions to the business experience unit by analyzing the trends across the historical datasets and shed lights on the optimal staffing level.

In addition to the time series analysis on predicting the baseline of incoming calls, we are interested in finding the relationship between the number of operators and the wait time. It helps us to identify the optimal staffing levels needed to meet certain service goals. I am going to use 3 models from queueing theory, ErlangC, ErlangB and ErlangA models to suggest the optimal level of staff needed.

In order to select the relevant features as inputs to the Erlang model. The scheme of the call center operations shown in figure 1.1 help us understanding the procedure of traffic flows. Initially, incoming calls form a single queue; customers wait for services. Statistically, they can be treated as a number of identically and independently ($i.i.d.$) distributed samples waiting for answering by N $i.i.d$ agents. Due to the impatient nature of human-beings, some of them might abandon the call during the queue and they might try again later or give up. Otherwise, costumers are automatically connected to an agent by system. During the active talk time, agents try to help customers to identify and solve their problems. Sometimes agents need to search the relevant information and have some discussion. In the meantime,

costumers are held in the system and waited to be answered. The conference time is the key component to solve issues. After call ended, staff would select the disposition group of calls, which is the type of the queries. We notice that it is an essential duty to understand the in-bound call procedure because some of parameters in the model, for instance, service time, are sensitive to the model accuracy. Therefore we need to make sure to use the right period of time in the whole scheme when implementing our model.

In this paper, we mainly focus on the dataset, applying the models to forecast and optimize the call center operations and report the summary of results and potential suggestions. The details of mathematical derivations are not in the scope of this paper.

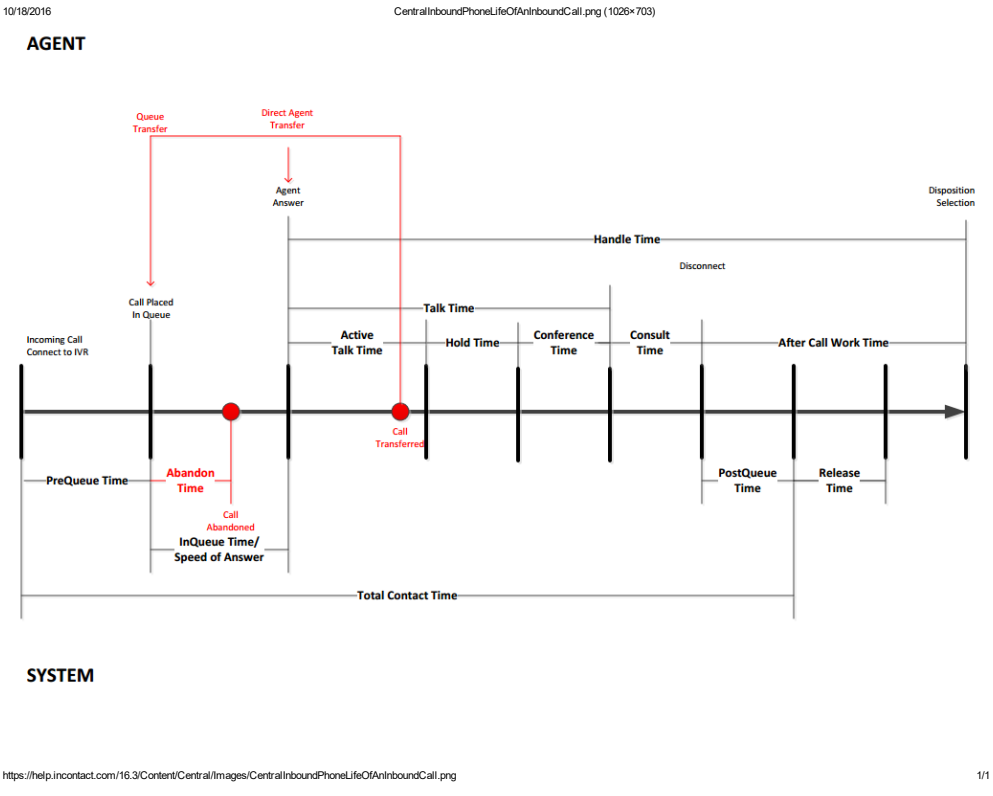


Figure 1.1: operational scheme of call center

CHAPTER 2

Literature Review

There are a large number of literatures, focusing on the time series forecasting of the call volume and model the in-bound call arrival process as a queueing system.

Walter Zucchini [16] gives us a glance of view of the procedure of modeling some basic time series forecasting using R. Rouba [13] Ibrahim explains the Holt-Winters exponential smoothing technique applying to the call center forecast in more details. Katleho [8] gives a comparative study between the seasonal ARIMA model and Holt-Winters exponential smoothing model.

The analysis between the number of operators and service time is a crucial part to improve the call center operations. In other words, we are interested in matching the service level (supply) to customers' request (demand). There are vast amount of papers on this topic and most of them use different models in queuing theory, which considers as an effective approach to model the traffic of calls. The ErlangC model is considered to be one of the simplest but effective models. Robbins and Mederios [15] conduct the most basic analysis using ErlangC model and give a detailed discussion on the accuracy on real call center scenario. D. Ekmekçiu [5] gives an emphasis on ErlangC model and conduct simulation model. Ger Koole [7] conducts the analysis using ErlangA model. Brown [9] give us a more advanced insights ErlangA model from a more statistical prospective. Gans Koole and Mandelbaum [10] gives us an comprehensive insight of call center operations from a more business prospective. In addition, it also provides a detailed discussion and comparison covering ErlangC, ErlangB and ErlangA model. Garnett, Mandelbaum and Reiman [11] present the model with an emphasis on customers' impatient behaviour. Robbins [14] describes the ErlangA model outperforms ErlangC model in detail. In our paper, based on these three methods,

we implement them our the City of Los Angeles real world dataset, comparing them and discussing results.

Furthermore, there are some useful online calculators that are easy for us to calculate the staffing level by implementing various models. One of the most widely used one is www.4callcenters.com. Another one is called 'CCOptim' [3], which allows us to do the straight forward calculations using ErlangA, ErlangB and ErlangC models; some visualzations are given allowing us to see the performance in a more straight forward prospect of view. Furthermore, there are some available R package like 'simmer' and 'queuecomputer' [1], which allows as to perform the simulation of queueing system.

CHAPTER 3

Dataset Description

The call center generates a large amount of records on the daily basis by the computer system called IVR and ACD mediate. At each time the call flows, these operating systems record some features including but not necessarily limit to the unique ID call number, handle time and the action taken.

There are two main components of the datasets, One contains the summary data for 3 years; the other are valuable raw datasets for 10 months, allowing as to extract the detailed information corresponding to the scheme in figure 1.1, say, service time, handle time, abandon time.

The dataset contains the summary data from 07/01/2014 to 09/13/2017. In this dataset, the feature which provides information of the total number of calls received is our main interest of variable to model the univariate time series model. This 3-year time series dataset is enough for us to make 1 month prediction using two univariate models, ARIMA model and exponential smoothing model.

Another one is the raw dataset contained the 45 detailed features of call center operations from 11/21/2016 to 8/31/2017. The table 3.1 contains a list of features. In particular, those include different periods of time presented in figure 1.1, contact start time, end time, abandons, etc.

In particular, 3 parameters we needed in ErlangC model can be extracted directly or easily computed from this dataset to help us to extract useful columns for modeling. For instance, the service time needed for ErlangC model, the figure 1.1 allows us to identify the service time, which can be extract from the feature named "handle.time", corresponding to the period from "Agent Answer" to the Dispersion selection. In the real scenario, some staff

might not complete the disposition selection immediately after the consulting ended. Hence there might be some outliers presented in the dataset should be identified and discarded. For the ErlangB model, the columns of interests are the same as the ErlangC, which can be directly acquired from our raw dataset. However, for the ErlangA model, one of the variables we are interested is the abandonment time, which is missing in our raw dataset. Hence we have to perform the simulation and experiments regarding the assumption of distribution of the abandonment rate.

Name of Variables	Type	Explanation
Date	date	actual date of contact
Contact ID	string	unique ID of contact
Direction	boolean	whether it is inbound or outbound
Handle.Time	time (hh:mm:ss)	duration of agents handling contacts
Contact.Start.Date.Time	date and time	time when contact starts

Table 3.1: Some important features of dataset

CHAPTER 4

Methodology

In this chapter, aiming to achieve the goal of forecasting the call volume of both baseline and peak seasons, I am going to use two univariate time series approaches, ARIMA (autoregressive integrated moving averages) model and Holt-Winters Smoothing technique. Finding the optimal staffing level of call center, we use the 3 different models in queueing theory in the following step.

4.1 Time Series Analysis

4.1.1 Introduction

Time series analysis is one of most useful tool applying to the business operations. It could explain aspects of historical patterns. More importantly, it can help us to predict the trend of a variable. In our call center operations problem, the seasonal patterns of call received, the expected number of baseline call volume can be predicted using time series approach.

In this chapter, I mainly introduce two methods of time series model to forecast the total call received in the peak season. The first one based on the ARIMA model and the second one is called *exponential smoothing* model. Both of these two methods aim to calculate the baseline of the call volumes, as well as that of peak season from 1^{st} Jan to 28^{th} Feb.

The dataset I used is the daily call received from 07/01/2014 to 9/13/2017, which includes the total number of calls received for more than 3 years, 790 days, which is enough for us to make the 1-month prediction.

4.1.2 ARIMA Model

The main ingredients of the ARIMA model contain the model identification, parameters estimation and diagnostic checking. There are 3 main parameters: p, d, q, which consists $ARIMA(p, d, q)$. p is the number of AR terms, d is the number of non-seasonal differences for stationarity and q is the number of lagged forecast errors in the prediction equation. Initially, we need to do the test to identify if we need to be differenced to make the time series (Y_t) *stationary*. In other words, to specify the value of d, if $d = 0$, we say the Y_t is stationary; otherwise is it non-stationary. Subsequently, we need to specify the ACF (autocorrelation) and PACF (partial autocorrelation) for AR (autoregressive) and MA (moving average) models, which helps us to identify values of p and q. Once the model is appropriately chosen and specified, we could fit the model to the historical time series data. After the model is fitted, the final step is to check whether is is make sense and use it to do the predictions and forecasts.

Given a time series data y_t, if it satisfies the following equation, we define it to be $ARIMA(p, d, q)$ model. The general forecasting equation is

$$\hat{y}_t = \mu + \phi_1 y_{t-1} + ... + \phi_p y_{t-p} - \phi_1 e_{t-1} ... - \theta_q e_{t-q} \tag{4.1}$$

4.1.2.1 Stationarity Test

In order to test the stationarity of the time series data of daily call received in the call center, I use the *unit root stationary test*. Dickey and Fuller [4] introduced a procedure to test whether a particular variable is stationary, which means it is a random walk and doesn't have specific pattern and trend of statistical aspects (mean, variance and autocorrelation etc.). The hypothesis testing of Augmented Dickey-fuller (ADF) test and test statistics are explained below:

Mathematical model $y_t = \alpha + \beta t + \gamma y_{t-1} + \delta_1 \Delta y_{t-1} + ... + \delta_p \Delta y_{t-p} + \epsilon_t$, where α is a constant, β is the coefficient on a time trend, Δ is the differencing operator, i.e., $\Delta y_t = y_t - y_{t-1}$, p is the lag order of the AR process and ϵ_t is a innovation process with mean zero. For the

null hypothesis, we state the time series is stationary. The alternative hypothesis is that the time series is non-stationary. The testing process can be formulated mathematically [6]:

$$H_0 : \gamma = 0$$

$$H_a : \gamma < 0$$

Test statistic:

$$DF_\tau = \frac{\hat{\gamma}}{SE\hat{\gamma}}.$$

where $\hat{\gamma}$ is the least squares estimate and $se(\hat{\gamma})$ is the usual standard error estimate. DF_τ is obtained and compared with the critical value of DF test. If the test statistics is less than the critical value, we reject H_0 and conclude that the time series in non-stationary. Also, we could compare the p-value at α significance level to draw the conclusion.

4.1.2.2 Model Order

The ACF plot, known as autocorrelation plot is a widely used graphical tool to determine the parameter q for the ARIMA model, the order of $MA(q)$. Another plot is called PACF plot (partical autocorrelation plot), explaining the autocorrelations which is not well explained by the former ACF plot. Similarly, it helps us to determine the order to $AR(p)$, parameter p.

The ACF plot is an bar chart of the correlation coefficients between the time series and lags of itself and the PACF plot is that of the partial correlation coefficients between the time series and lags of itself. The sharp reduction of the magnitude in autocorrelation and partial autocorrelation helps us to determine the order. For instance, in a PACF plot, if the magnitude of second lag deceases significantly, then we might conclude that the $q = q$. Furthermore, the higher order autocorrelations are effectively explained by the lag-1 autocorrelation and $p = 1$.

4.1.3 Exponential Smoothing Model - Additive Seasonal Model

In this section, we introduce one of the most popular Holt-Winters exponential smoothing model called additive seasonal model, which is commonly used for the call center operations. This forecasting technique is basically constructed from the exponentially weighted average of the historical observations. We choose this time forecasting technique because it might give us a better forecasting result than the result of ARIMA model because it considers seasonal effects.

Furthermore, there are two methods of seasonal Holt-Winters. One is additive method and the other is multiplicative method. For the additive models, the amplitude of seasonal variation is independent of the level. For the multiplicative model, the level and seasonality are connected. The time series data can be modelled below [12],

$$y_t = b_1 + b_2 t + S_t + \epsilon_t$$

where b_1 is the signal that cannot be explained by the trend and seasonal effect, which is also called *permanent component*, b_2 is the linear trend component, S_t is the additive seasonal effect and ϵ_t is the random error component.

4.1.3.1 Updating Parameters

There are 3 smoothing parameters namely α, β and γ, which represents for the deseasonalized level, trends and seasonal effect correspondingly. Let $\bar{R}_t$ be the estimate of the deseasonalized level, $\bar{G}_t$ be the estimate of trend and $\bar{S}_t$ be the estimate of seasonal effect. The estimates of $\bar{R}_t$, $\bar{G}_t$ and $\bar{S}_t$ are listed below:

- Level smoothing

$$\bar{R}_t = \alpha(y_t - \bar{S}_{t-L}) + (1-\alpha)(\bar{R}_{t-1} + \bar{G}_{t-1})$$

  where L is the length of the seasonal periods.

- Trend smoothing

$$\bar{G}_t = \beta(\bar{S}_t - \bar{S}_{t-1}) + (1-\beta)\bar{G}_{t-1}$$

- Seasonal effect smoothing

$$\bar{S}_t = \gamma(y_t - \bar{S}_t) + (1-\gamma)\bar{S}_{t-L}$$

Note that values of α, β and γ are 3 smoothing constant which are between 0 and 1. Now we could set up the value of forecast for the next period:

$$y_t = \bar{R}_{t-1} + \bar{G}_{t-1} + \bar{S}_{t-L}$$

.

4.2 Queuing Theory

Queuing Theory is a mathematical tool to study the long waiting lines. It has intensive applications in the field of operational analysis. In this section, we focus on the analysis of ErlangC, ErlangB and ErlangA to figure out the number of servers needed to achieve the target service level.

4.2.1 ErlangC model

The ErlangC model, also known as $M/M/N$ queue, is one of the simplest and widely used model on the multi-server system. There are numbers of assumptions in this model. Firstly, the calls are not served immediately; they are all put into a wait queue and wait for service. Secondly, this model assumes that the call arrival is a Possion process at a average rate of λ. Thirdly, the service duration are exponentially distributed with a mean service time of $\frac{1}{\mu}$. In addition, we assume that both of customers and servers are independently and identically distributed. Finally, we assume the impatience and abandonment behaviours of costumers

are disregarded, which means people wait in the queue and never hang up the phone before the response.

After the assumption of ErlangC model is specified, we introduce parameters of the model. All of these features can be collected or easily calculated from our dataset the call center provided. There are 3 parameters needed to input into this model. The call arrival rate (λ), call duration (T_s), and the number of agents (N). We record the average rate as the number of calls received per hour and the call duration is the average call duration per hour. We should keep the time units the same when specifying call arrival rate, λ and duration, T_s.

$$\lambda = \text{the number calls}/3600 \tag{4.2}$$

The traffic intensity (u) can be calculated by,

$$u = \lambda T_s. \tag{4.3}$$

The agent occupancy (ρ) is then calculated by dividing the traffic intensity by the number of agent presented,

$$\rho = u/N \tag{4.4}$$

where $\rho \in (0, 1)$. The next crucial step is to calculate the Erlang-C formula, which is also known as the probability of waiting [15],

$$P(\text{wait} > 0) = \frac{\frac{u^N}{N!}}{\frac{u^N}{N!} + (1-\rho)\sum_{k=0}^{N-1} \frac{u^k}{k!}}. \tag{4.5}$$

Then the average speed of answer (ASA) can be calculated, which is the average waiting time for a call can be calculated easily from $P(\text{wait}) > 0$.

$$ASA = \frac{P(\text{wait} > 0)T_s}{N(1-\rho)}. \tag{4.6}$$

Finally, the service level $W(t)$, the probability of a call will be answered with in a target wait time (t), can be computed.

$$W(t) = P(\text{wait time} \leq t) = 1 - P(\text{wait} > 0)e^{-(N-u)\frac{t}{T_s}}. \tag{4.7}$$

Now we end up with finding the relationship between the number of agents (N) and the probability of wait time within a time limit ($W(t)$), which gives us an insight of the optimal service level to capture the phone calls.

4.2.2 ErlangB model

The ErlangB model ($M/M/n/n$) helps us to compute the probability of *blocking calls* (P_b), which refers to the calls that have not been completed due to the intensive traffic. Then it works out the relationship between P_b and service lines needed during the business hour traffic given the number of hours of call traffic.

The assumptions of this model are similar to ErlangC model. The arrival process is a *Possion process*, the duration of the time that a user occupied the channel is *exponentially distributed* and infinitely number of arrivals. One main difference which should be emphasized is that comparing with ErlangC model, instead of assuming no abandonment behaviour, we assume that for users in the service line, there are no queues and users get responses immediately if there are channels available. Otherwise, the requesting user is blocked without access and they could try again later.

Using the same notation as in ErlangC in the previous section, we denote μ as the traffic intensity and N is the number of servers in channels. The call blocking probability P_b is calculated by the following equation,

$$P_b = \frac{u^N/N!}{\sum_{i=0}^{N} \frac{u^i}{i!}}. \tag{4.8}$$

4.2.3 ErlangA model

In previous two sections, we discuss the ErlangC and ErlangB models. In ErlangC, one of the most important assumption is that there is no abandonment rate. In ErlangB, it assumes that if calls were not answered immediately, the call is blocked. It is obvious that these assumptions are naive and problematic. The impatient nature of human sometimes cause the abandonment of calls and others might wait in the queue for a couple of minutes. One trade off strategies between ErlangB and ErlangC is to consider the abandonment behaviour. Therefore we consider another model, called ErlangA ($M/M/N+M$), allowing busy signal and abandonment to be considered.

Keeping all other assumptions the same, the only difference between ErlangA and ErlangC is that the former allows abandonment. In addition to ErlangC model ($M/M/N$), we enrich the model to ($M/M/N+M$) by considering the abandonment rate, which assumes to be exponentially distributed.

Keeping the notation of ErlangA the same as before, we use λ as the arrival rate, μ as the service rate (the reciprocal of service time recorded hourly), N representing the number of servers. The model and exact formulae [2] are given below.

Recall that the patience time is iid and exponentially distributed with parameter θ, which can be expressed as a function of H:

$$H(x) = \frac{1}{\theta}(1 - e^{-\theta x}). \tag{4.9}$$

Define the *incomplete Gamma function*, which is required for as to calculate the probability of waiting,

$$\gamma(x, y) \triangleq \int_0^y t^{x-1} e^{-t} dt, x > 0, y \geq 0. \tag{4.10}$$

Then the building block can be expressed by,

$$J = \frac{e^{\lambda/\theta}}{\theta}(\frac{\theta}{\lambda})^{\frac{N\mu}{\theta}} \gamma(\frac{N\mu}{\theta}, \frac{\lambda}{\theta}). \tag{4.11}$$

Define ϵ,

$$\epsilon = \frac{\sum_{j=0}^{N-1} \frac{1}{j}(\frac{\lambda}{\mu})^j}{\frac{1}{(N-1)!}(\frac{\lambda}{\mu})^{N-1}}. \tag{4.12}$$

Finally, we can calculate the probability of waiting,

$$P(\text{Wait} > 0) = \frac{\lambda J}{\epsilon + \lambda J}(1 - \theta). \tag{4.13}$$

CHAPTER 5

Results and Comparison

5.1 Results and Discussion of Time Series Analysis

5.1.1 Results of ARIMA Model

5.1.1.1 Determination of Optimal Model

Following the section 4.1.2.1, we apply the testing procedure on the total daily call received, we use the function adf.test() in tseries package in R to find the following result. Dickey-Fuller $= -6.0412$, lag order $= 9$ and $p - value = 0.01$. We conclude that this time series is stationary at 0.05 significance level. From figure 5.1, we can see that the fluctuations, means are roughly constant through out the time. Intuitively, it coincides the ADF test results we test here. Therefore, we set the optimal value of d to 0.

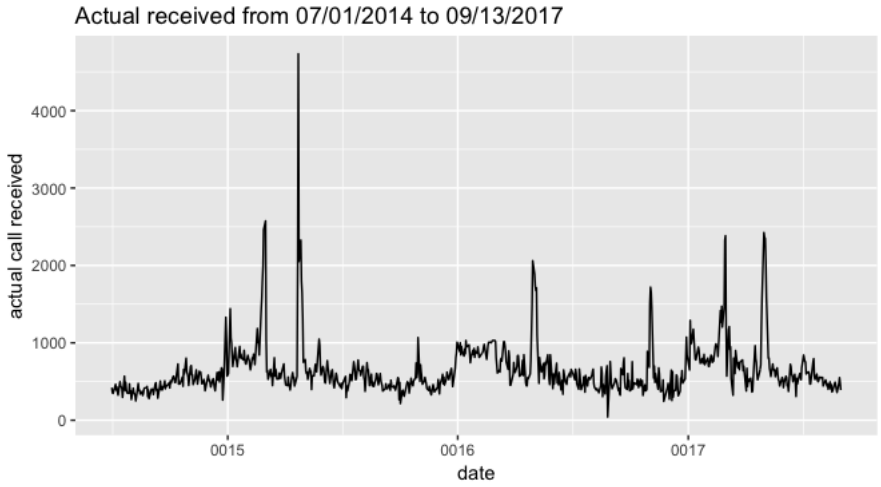


Figure 5.1: plot of historical time series data of actual call received

Subsequently, we compute the model order by the method in section 4.1.2.2. From figure

5.2, the ACF plot shows that the first 10 lags are not ignorable, it might be hard to draw the conclusion on parameter q from this single graph. However, from the PACF plot, the partial autocorrelation of the second lag is significantly reduced. Combine the features from these two graphs, we might conclude that $p = 1$ and $q = 1$ because the propagation at lag 1 contributes to the autocorrelation at lag 2, which is verified by the second graph. To conclude, we choose the fitted model to be $ARIMA(1, 0, 1)$.

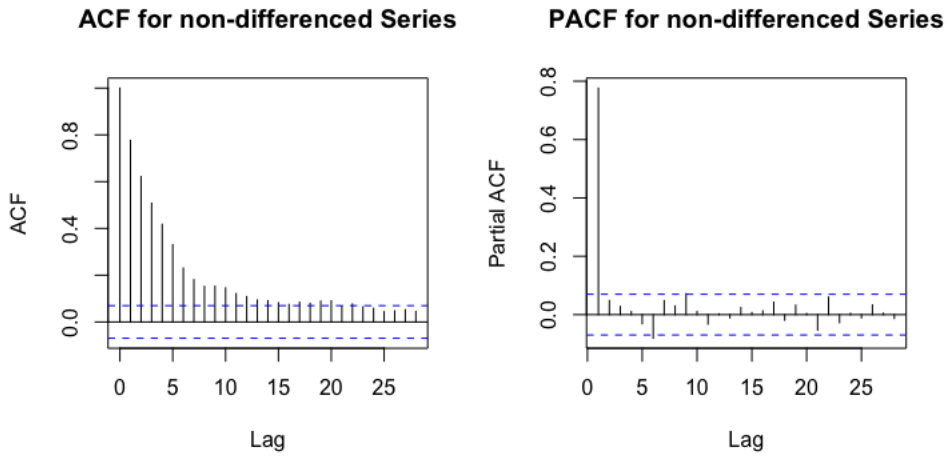


Figure 5.2: plots to determine the parameters p and q

5.1.1.2 Model Validation

In order to validate the model, we need to apply the diagnostic checking on our $ARIMA(1, 0, 1)$ model. From figure 5.3, despite a few large values, we notice that residuals are roughly normally distributed. We could conclude that this time series model is generally validated and could be used for future forecasting of call volume of the baseline and peak season.

5.1.1.3 Forecasting

Once the model has been specified, we could use this method to predict the future call volume. The table 5.4a predicts the baseline of call volume in August, which overestimates the actual call received by 188 on average. However, for the peak season prediction in Jan,

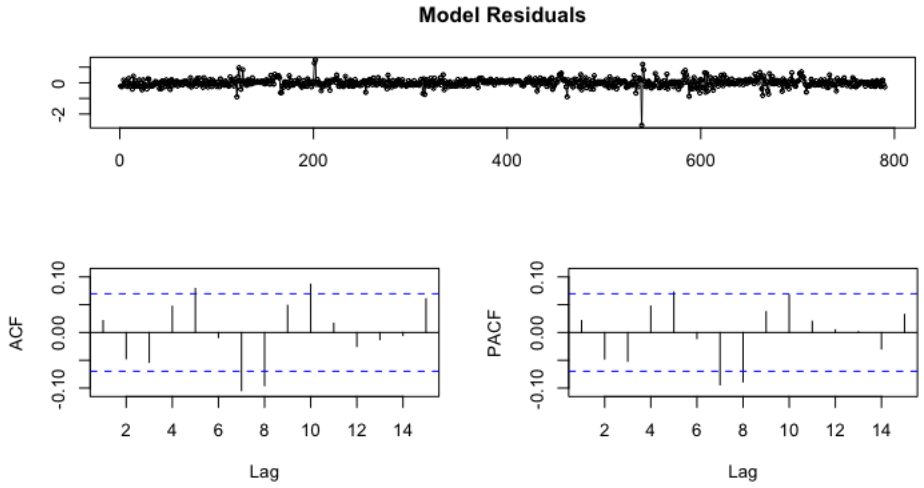


Figure 5.3: diagnostic checking using residuals plot

this model underestimates the call volume by 196.

From figure 5.5, the blue line represents the forecast of the call volume. From the historical data, we notice that the total call volume is fluctuated on daily bases. It is very unlikely that the call volume stays unchanged in the future. Furthermore, it also gives a wide range of prediction interval as we notice from table 5.4 and figure 5.5. This model seems to be a simple and naive one. One important fact is that we haven't included the seasonality in this model. Some other model such as exponential smoothing will be introduced in the subsequent section aiming to improve the prediction results.

DATE	ACTUAL CALLS RECEIVED	POINT ESTIMATE	DIFFERENCE	Lower 90%	Higher 90%	Lower 95%	Higher 95%
8/1/17	539	576	37	202	950	131	1022
8/2/17	554	595	41	122	1068	31	1159
8/3/17	443	610	167	86	1133	-14	1234
8/4/17	462	621	159	70	1173	-36	1279
8/7/17	521	630	109	62	1198	-47	1307
8/8/17	491	637	146	59	1214	-51	1325
8/9/17	442	642	200	59	1225	-53	1337
8/10/17	469	646	177	60	1232	-52	1345
8/11/17	400	649	249	61	1237	-52	1350
8/14/17	496	652	156	62	1241	-51	1354
8/15/17	405	654	249	63	1244	-50	1357
8/16/17	460	655	195	64	1246	-49	1359
8/17/17	366	656	290	65	1247	-48	1360
8/18/17	377	657	280	66	1248	-47	1361
8/21/17	468	658	190	67	1249	-47	1362
8/22/17	492	658	166	67	1249	-46	1363
8/23/17	449	659	210	68	1250	-46	1363
8/24/17	400	659	259	68	1250	-45	1363
8/25/17	359	659	300	68	1250	-45	1364
8/28/17	478	659	181	68	1251	-45	1364
8/29/17	550	660	110	68	1251	-45	1364
8/30/17	477	660	183	68	1251	-45	1364
8/31/17	390	660	270	69	1251	-45	1364
AVG	456	644	188				

(a) baseline prediction

DATE	ACTUAL CALLS RECEIVED	POINT ESTIMATE	DIFFERENCE	Lower 90%	Higher 90%	Lower 95%	Higher 95%
1/3/17	650	813	163	442	1185	370	1256
1/4/17	1290	774	-516	324	1224	238	1310
1/5/17	985	744	-241	250	1239	155	1334
1/6/17	1001	721	-280	200	1242	100	1342
1/9/17	1176	702	-474	165	1239	62	1342
1/10/17	1037	687	-350	141	1234	36	1339
1/11/17	901	675	-226	123	1228	17	1334
1/12/17	788	666	-122	110	1223	3	1329
1/13/17	779	659	-120	100	1217	-7	1324
1/17/17	945	653	-292	93	1213	-15	1320
1/18/17	898	648	-250	87	1209	-21	1317
1/19/17	799	644	-155	83	1206	-25	1314
1/20/17	735	641	-94	79	1204	-28	1311
1/23/17	799	639	-160	77	1201	-31	1309
1/24/17	739	637	-102	75	1200	-33	1308
1/25/17	738	636	-102	73	1198	-35	1306
1/26/17	851	635	-216	72	1197	-36	1305
1/27/17	718	634	-84	71	1196	-37	1304
1/30/17	774	633	-141	70	1196	-38	1303
1/31/17	787	632	-155	70	1195	-38	1303
AVERAGE	870	674	-196				

(b) peak season prediction

Figure 5.4: ARIMA model forecasting

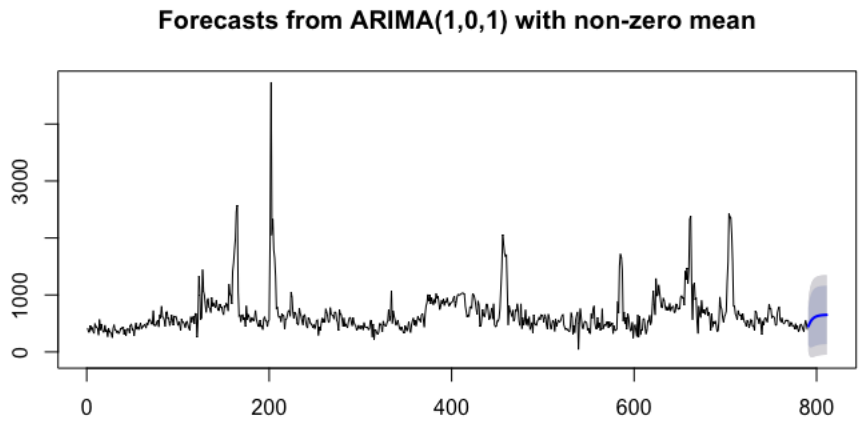


Figure 5.5: forecasting using ARIMA(1, 0, 1)

5.1.2 Results of Exponential Smoothing Model

5.1.2.1 Determination of Optimal Model

From figure 5.1, we notice that the trend and seasonality appear to be independent. Therefore we choose the Holt-Winters exponential smoothing with trend and additive seasonal component here.

5.1.2.2 Forecasting

The $HoltWinters()$ function in 'ts' library in R allowed us to find these 3 parameters by minimizing the mean square prediction errors. The dataset that I am going to use are the same as that we use in the ARIMA model for the purpose of comparing the predictive accuracies of these two time series forecasting techniques. We find the $\alpha = 0.7601294$, $\beta = 0$ and $\gamma = 0.204224$.

From figure 5.6, the red line at after the dotted vertical line is the prediction of call volume on Jan 2017. The table 5.10a gives us the baseline forecasting and 5.10b shows the peak season forecasting using the seasonal additive model for the exactly the same period prediction of ARIMA model. We notice that this model gives us a better prediction. For

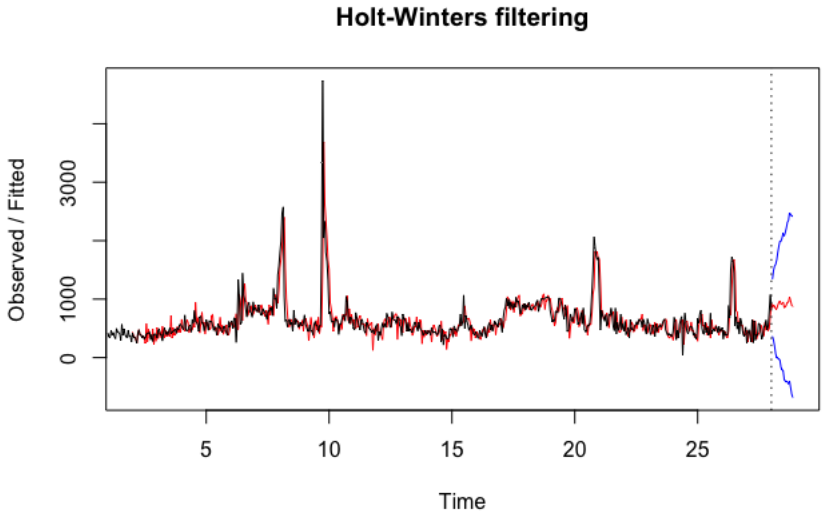


Figure 5.6: prediction of seasonal additive model on Jan 2017

the call volume in the peak season, the average difference is reduced to 40 and to 68 for the baseline.

5.1.3 Discussion of Times Series Analysis

In this particular case, comparing the last row of table 5.4a and 5.10a, we could see that Holt-Winters method performs significantly better than the ARIMA model. The difference of prediction and the actual value is 68 for HW method and that of ARIMA method is 188. The reason might be that the HW method considers the effects seasonality.

5.2 Results of Queueing Theory

5.2.1 Results of ErlangC Model

The table 5.1 gives us the column used from the raw dataset to address 3 main parameters.

Following by this set up and the raw dataset given, we are able to find the relationship between the number of agents presents and the proportional of calls captured within certain time limit.

DATE	ACTUAL CALLS RECEIVED	POINT ESTIMATE HW	DIFFERENCE HW
8/1/17	539	589	50
8/2/17	554	549	-5
8/3/17	443	562	119
8/4/17	462	629	167
8/7/17	521	523	2
8/8/17	491	559	68
8/9/17	442	538	96
8/10/17	469	580	111
8/11/17	400	591	191
8/14/17	496	641	145
8/15/17	405	481	76
8/16/17	460	494	34
8/17/17	366	517	151
8/18/17	377	514	137
8/21/17	468	496	28
8/22/17	492	455	-37
8/23/17	449	441	-8
8/24/17	400	497	97
8/25/17	359	500	141
8/28/17	478	441	-37
8/29/17	550	441	-109
8/30/17	477	502	25
8/31/17	390	522	132
AVG	456	524	68

(a) baseline prediction

DATE	ACTUAL CALLS RECEIVED	POINT ESTIMATE HW	DIFFERENCE HW
1/3/17	650	853	203
1/4/17	1290	893	-397
1/5/17	985	880	-105
1/6/17	1001	831	-170
1/9/17	1176	836	-340
1/10/17	1037	894	-143
1/11/17	901	942	41
1/12/17	788	973	185
1/13/17	779	910	131
1/17/17	945	918	-27
1/18/17	898	952	54
1/19/17	799	856	57
1/20/17	735	867	132
1/23/17	799	899	100
1/24/17	739	952	213
1/25/17	738	948	210
1/26/17	851	1037	186
1/27/17	718	970	252
1/30/17	774	898	124
1/31/17	787	872	85
AVERAGE	870	909	40

(b) peak season prediction

Figure 5.7: Seasonal additive Holt-Winters model forecasting

Variable of interest	Column used from dataset
Number of agents (N)	CSS ON PHONES
Call duration (T_s)	Handle.Time
call arrival rate (λ)	# Contact.Start.Date.Time/3600

Table 5.1: Parameters extracted from raw dataset

From table 5.8, 2^{st}, 3^{nd} and 5^{th} columns (highlighted by dark orange) are 3 input parameters which correspond the 3 rows in table 5.1. Setting the wait time to 5 minutes, the intensity, Occupy. Prob wait, ASA and Wait $<$ tt are calculated by equation 4.3, 4.4, 4.5, 4.6 and 4.7 correspondingly.

The traffic time are mainly cumulated during 9 : 00 AM to 11 : 00 AM. We could see that 29 agents capture above 90% of calls within 5 minutes. On 27^{th} April, 32 agents are not enough to capture the traffic from 9 : 00 AM to 10 : 00 AM.

The table 5.9 shows the number of servers and the service level from 10 AM to 11 AM on 28^{th} Feb. In addition, the service levels and delay probabilities and the average waiting times against the number of servers are shown in figure 5.10.

2/28/17								
Date & Time	The number of calls	Average of Handle.Time.Sec	Intensity	Agents	Occup.	Prob wait	ASA	Wait < tt
7:00 AM	6	231.00	0.4	29.00	1%	0%	0	100%
8:00 AM	240	306.80	20.5	29.00	71%	5%	2	100%
9:00 AM	255	350.82	24.8	29.00	86%	32%	27	99%
10:00 AM	273	350.71	26.6	29.00	92%	55%	80	93%
11:00 AM	244	266.49	18.1	29.00	62%	1%	0	100%
12:00 PM	238	272.71	18.0	29.00	62%	1%	0	100%
1:00 PM	224	302.05	18.8	29.00	65%	2%	1	100%
2:00 PM	226	283.21	17.8	29.00	61%	1%	0	100%
3:00 PM	216	220.85	13.3	29.00	46%	0%	0	100%
4:00 PM	173	278.66	13.4	29.00	46%	0%	0	100%
5:00 PM	1	56.00	0.0	29.00	0%	0%	0	100%
4/27/17								
1:00 PM	234	311.21	20.2	23.81	85%	32%	27	99%
2:00 PM	238	290.02	19.2	23.81	81%	22%	13	100%
3:00 PM	170	208.78	9.9	23.81	41%	0%	0	100%
6:00 AM	5	442.74	0.6	23.81	3%	0%	0	100%
7:00 AM	228	368.60	23.3	23.81	98%	85%	477	50%
8:00 AM	231	324.62	20.8	23.81	87%	40%	41	98%
9:00 AM	245	364.34	24.8	23.81	104%	120%	-551	-132%
10:00 AM	231	254.38	16.3	23.81	69%	5%	2	100%
11:00 AM	239	316.36	21.0	23.81	88%	43%	45	98%
12:00 PM	214	256.70	15.3	23.81	64%	3%	1	100%

Figure 5.8: Results of 2 sample days by ErlangC model

5.2.2 Results of ErlangB Model

From the results displayed by ErlangC model, we could see that the busy hour traffic is especially during the morning mainly from 9 : 00 AM to 12 : 00 PM in the peak season (January and February).

In table 5.2, results are shown by using ErlangB model. Setting $P_b = 0.05$, which means there are 5 calls are blocked out of 100 calls, I compute the number of servers needed (N) corresponding the last column in table 5.2. The first three columns using the same data in table 5.8. During the peak hours of the day, especially from $9:00AM$ to $11:00AM$, the number of channels are ideally needed to be increased to 31 or 32. The plot of blocking probabilities against the number of servers is shown in figure 5.11.

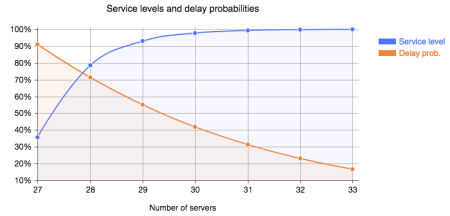
5.2.3 Results of ErlangA Model

Now we implement the ErlangA theory on the information 28^{th} Feb. 2018. The calculation is performed by the online calculator [3]. We notice that we have no features of records on the average patience time, which indicates the average time duration of a costumer staying in the queue. For the simplicity, we choose the threshold to be 10 minutes in our calculations. The results of minimum number of operators needed to achieve the target (95% of calls answered

Minimum number of servers needed is **30** for service level target of **95%**.

Servers	Service Level (%)	Delay (%)	Avg Wait (minute)
27	35.64	90.97	13.15
28	78.56	71.29	2.97
29	92.96	55.06	1.34
30	**97.72**	**41.87**	**0.72**
31	99.28	31.33	0.42
32	99.77	23.05	0.25
33	99.93	16.67	0.15

Figure 5.9: tables of servers against service level



(a) service levels and delayed probabilities vs servers

(b) average serving time vs servers

Figure 5.10: plots of delay and wait time against the number of servers from $10:00$ AM to $11:00$ AM on 28^{th} Feb. using ErlangC

within 5 minutes) are shown in the last column of table 5.2.

In particular, some additional information of ErlangA model are shown in figure 5.13a and figure 5.13. Figure 5.13a allows us to see the relationship between the number of servers needed and the service level. For example, if we want 90% of calls are answered within 5 minutes, 20 servers are needed. Furthermore, the visualizations of percentage of delay, abandonment and the average wait time against servers are shown in details in figure 5.13.

Time	No. of calls	Avg Handle.Time.Sec	P_b	N (ErlangB)	N (ErlangC)
7:00 AM	6	231.00	0.05	2.39	2
8:00 AM	240	306.80	0.05	26.45	17
9:00 AM	255	350.82	0.05	30.85	20
10:00 AM	273	350.71	0.05	32.60	22
11:00 AM	244	266.49	0.05	23.06	15
12:00 PM	238	272.71	0.05	23.03	15
1:00 PM	224	302.05	0.05	24.79	16
2:00 PM	226	283.21	0.05	23.78	15
3:00 PM	216	220.85	0.05	18.25	11
4:00 PM	173	278.66	0.05	18.39	12
5:00 PM	1	56.00	0.05	1.02	1

Table 5.2: calculation of number of servers needed for on 02/28/2017

5.2.4 Discussion of Queueing Theory

As we have the actual raw dataset for the call center in the LA city, one simple approach of evaluating the accuracy and performance of the model is to compare the results with the actual data. From the real dataset, we could see the real situation of the call center. The number of operators presented are 16 and 38.8% of calls are abandoned on 28^{th}, February, 2017.

These 3 models are designed for call centers regarding the number of in-bounded call only. It is obvious that ErlangC model might overestimate the number of agents and underestimate the service level because it assumes no abandonment rate, which is unrealistic in today's call center operations. The average abandonment is 10.2%, which is not a negligible amount. On the other hand, the ErlangA model tends to make the overly optimistic predictions. We suggest that we combine the results of ErlangC and ErlangA, treating them as a confidence interval to determine the optimal staff level.

Furthermore, based on the model we used to calculate the optimal number of operators

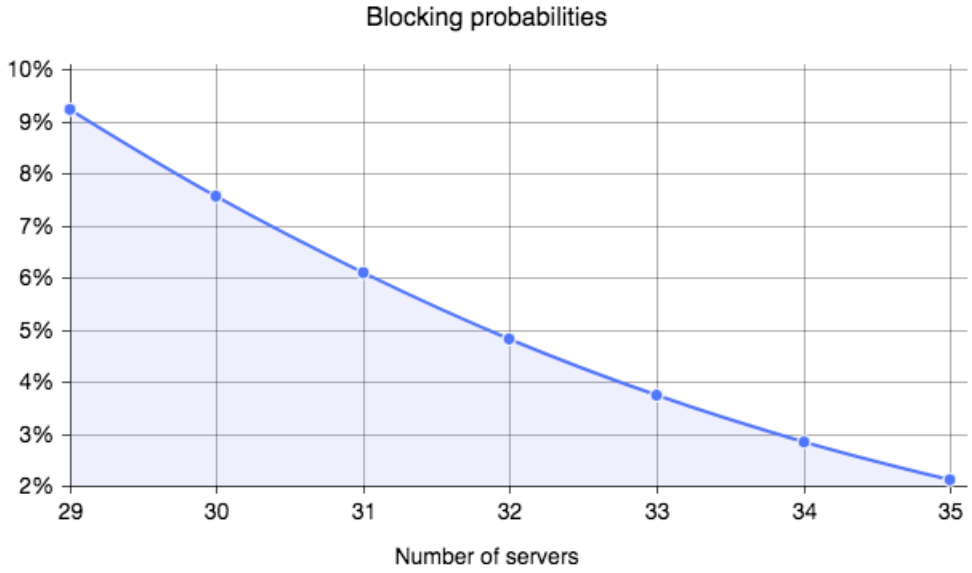


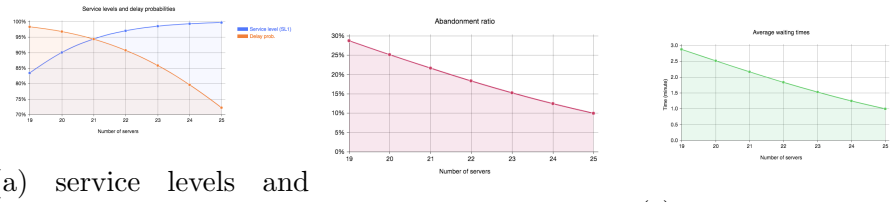
Figure 5.11: plot of blocking probabilities against the number of servers

needed, in order to further make the thoughtful and wise decisions. We should pay also attention to the costs of operators and the staffing cost to find a trade off strategy.

Minimum servers needed is **22** for target of **95%** with service level formula **SL1**.

Servers	Service Level (%) SL1	Delay (%)	Abandonment (%)	Avg Wait (minute)
19	83.41	98.31	28.71	2.87
20	90.02	96.78	25.11	2.51
21	94.38	94.33	21.63	2.16
22	**97.02**	**90.72**	**18.32**	**1.83**
23	98.51	85.81	15.23	1.52
24	99.29	79.58	12.43	1.24
25	99.68	72.21	9.95	0.99

Figure 5.12: relationship between the number of servers and



(a) service levels and delayed probabilities vs servers

(b) abandonment ratio vs servers

(c) average wait times vs servers

Figure 5.13: plots of delay, abandonment and average wait time against the number of servers from 10 : 00 AM to 11 : 00 AM on 28^{th} Feb. using erlangA

CHAPTER 6

Improvements and Further Research

In further research, I am interested in finding some other methods to optimize the call center operations. For instance, if we were to introduce the methods of web-chat and posting frequently asking questions and answers on our website, the traffic of answering the phone call can be potentially reduced.

In a more advanced setting and analysis, some latest machine learning techniques can be implemented. One of the most direct approaches is the speech recognition. For instance, we could use the speech-to-text engine to record and count the most frequent words and try to identify common problems. The deep learning method can be used to improve the accuracy of the speech recognition. Subsequently, answers appeared on the official website would become more targeted, inclining to reduce the traffic of the in-bounded calls.

CHAPTER 7

Conclusion

Call center in LA city confounds the traffic during January, February every year. We are interested in solving problems that how many operators needed to increase to solve the traffic issues. Furthermore, the baseline of the call volume are their another main interests to set the optimal number of staff and operators.

In this paper, we mainly help the call center to answer these two most important questions. We determine the baseline of the call center staffing level and the relationship between the number of agents and the wait time. The time series analysis and queueing theory approach give us useful insights supporting call center to make wise decisions. We use the Holt-Winters exponential smoothing method to compute relatively accurate forecast of the baseline call volume. Combining 3 different models in queueing theory, we have clear insights of the relationship between the number of operators and the wait time.

Based on the statistical approach to the call center operations improvements. Hopefully these conclusions could help them to raise the revenues and improve the services in the future.

REFERENCES

[1] A. Ebert, P. Wu, K. Mengersen and F. Ruggeri (2017), Computationally Efficient Simulation of Queues: The R Package queuecomputer.

[2] A. Mandelbaum, and S. Zeltyn, (2009), The M/M/n+G Queue: Summary of Performance Measures, Technical Note, Technion, Israel Institute of Technology.

[3] CCOptim, available at: https://www-ens.iro.umontreal.ca/ chanwyea/erlang/erlangA.-html.

[4] D. A. Dickey and W. A. Fuller, (1979), Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association* (74): 427431.

[5] D. Ekmekçiu, (2017), Queuing Models - A Call Center Case, *International Journal of Science and Research (IJSR)*, 6(2).

[6] E. Zivot, Unit Root Test, available at: https://faculty.washington.edu/ezivot/econ584/-notes/unitroot.pdf.

[7] G. Koole and A. Mandelbaum, (2001), Queueing Models of Call Centers An Introduction.

[8] K. D. Makatjane, M. Ntebogang and D. Moroke, (2016), Comparative study of Holt-Winters Triple Exponential Smoothing and Seasonal ARIMA: Forecasting Short Term Seasonal car sales in South Africa, *Risk governance control: financial markets & institutions*, 6(1).

[9] L. Brown et al., (2005), Statistical Analysis of a Telephone Call center., *Journal of the American Statistical Association*, 100(469).

[10] N. Gans, G. Koole and A. Mandelbaum, (2003), Telephone Call Centers: Tutorial, Review, and Research Prospects. Manufacturing Service Operations Management 5(2):79-141. available at: https://doi.org/10.1287/msom.5.2.79.16071.

[11] O. Garnett and A. Mandelbaum and M. Reiman, (2002), Designing a Call Center with Impatient Customers, Manufacturing and Service Operations Management 4 (3) : 208–227.

[12] P. S. Kalekar, *Time series Forecasting using Holt-Winters Exponential Smoothing*, available at: https://labs.omniti.com/people/jesus/papers/holtwinters.pdf.

[13] R. Ibrahim and P. L'Ecuyer, *Forecasting call center arrivals: A comparative study.*

[14] T. R. Robbins, (2016), Evaluating the fit of the ErlangA model in high traffic call centers, *Proceedings of the 2016 Winter Simulation Conference.*

[15] T.R. Robbins, D. J. Medeiros and T. P. Harrison (2010), Does the Erlang C model fit in real call centers? *Proceedings of the 2010 Winter Simulation Conference*.

[16] W. Zucchini, O. Nenadić, *Times Series Analysis with R - Part I*.