Федеральное государственное автономное образовательное учреждение

высшего образования «Национальный исследовательский университет   
«Высшая школа экономики»

###### **Факультет компьютерных наук**

###### **КУРСОВАЯ РАБОТА**

Изменение операционных параметров системы при увеличении нагрузки

Changing of System Operating Parameters with Increasing Load

по направлению подготовки 01.04.02 Прикладная математика и информатика

образовательная программа «Науки о данных»

Студент группы мНОД20\_ИССА\_\_ \_Юрьева Ирина Олеговна\_\_\_\_\_\_\_\_

(Ф.И.О.)

Руководитель КР \_профессор\_\_\_\_ \_Пьяных Олег Станиславович\_\_\_\_

(должность, звание, Ф.И.О.)

Москва, 2021

**Abstract**

In a queueing system with an increasing load, this load cannot increase forever without compromising the efficiency of the system. Based on the open dataset of the call center of one anonymous bank, this work investigated the behavior of the efficiency of queue systems in several parameters, such as the variability of the average waiting time in the queue, the load on operators, and the number of customers who dropped a call without waiting for a response. As a result, we were able to observe how exactly the system stops coping with the load when it becomes too high. We have also found that the formulas of the classical queue theory do not fit the system represented by this call center.

Table of Contents

[1. Introduction 1](#_Toc74693215)

[1.1 Related Works 1](#_Toc74693216)

[2.1 About call center dataset. 2](#_Toc74693217)

[2.2 Data preparation 3](#_Toc74693218)

[2.3 Visualizing results 3](#_Toc74693219)

[3. Results 4](#_Toc74693220)

[3.1 Waiting time metrics 4](#_Toc74693221)

[3.2 Service quality 5](#_Toc74693222)

[4. Testing the Classical Queuing Theory 6](#_Toc74693223)

[5. Conclusion 7](#_Toc74693224)

[6. References 9](#_Toc74693225)

# 

# 1. Introduction

In the modern world, there are many service systems with queues, such as call-centers, taxis, delivery, or regular queues in organizations. Queue theory investigates them and develops specific algorithms to improve the performance of these systems, such as the number of customers served and the quality of service using the least amount of cost (using load prediction and other methods). But it seems impossible to develop a perfect service system to work equally well at every moment of time. The load on the system cannot increase indefinitely, without affecting the quality of the system. While the classical queueing theory predicts infinitely high waits as the system approaches to 100% utilization, this is not enough for many practical applications, when one needs to know how exactly the quality of the system depends on the load. The purpose of our work was to study this dependency with the help of real data.

## Related Work

The research on queueing systems has been developed over a long period of time, starting at the early 20th century, when the queuing theory was originally created in the works of A. K. Erlang (Erlang, The Theory of Probabilities and Telephone Conversations, 1909) (Erlang, Telephone waiting times, 1920). Since then, queuing theory has explored different types of queues, based on Kendall's Notation for Queues (Kendall, 1953), depending on their different distributions of arrivals, services, different numbers of servers, limits, and priorities.

To better understand the features of different queues, researchers consider queueing models based on the described queue types that are closest to real data and perform analysis based on these models, as well as develop various applications to facilitate the work of service system managers (Mohammad Delasay A. I., 2014)(Saltzman, 2005). For example, the study in (Mohammad Delasay, 2016) developed a queuing model where service time depends on the load of the system and "tiredness" of the system to demonstrate the advantages of this model over models which ignore adaptive server behavior. Simulation models of call centers were also developed in (Robbins, 2006) (Pichitlamken, 2004). These studies help to understand how queues behave in certain situations, but in my opinion, multiple studies on real data provide a more useful and accurate understanding of the nature of different queues.

As part of the study of queue theory, there are also quite detailed studies of real service systems. In (Noah Gans N. L., 2010) were studied the operational heterogeneity of call center agents to understand how to increase average service by developing a better strategy for distributing customers across services. In (Lawrence D Brown, 2005), the authors present a large study of one detailed dataset, based on the data our work was written. However, all these studies describe datasets from the point of view of the classical queuing theory, but they do not pay much attention to the efficiency of the system and what happens to the system under increased load. We note that there are not many studies based on real data, since not all organizations collect data suitable for analysis using queue theory, and not all organization ready to share the data.

With an extensive search for articles, we managed to find only one study (Ludwig Kuntz, 2011) on the topic of system overload, while preparing our work. Regarding the theory of queues, this article proved that the impact of load on the quality of service is nonlinear.

1. Methods and materials

|  |
| --- |
| Diagram  Description automatically generated |
| Figure 1 Event history of calls (units are calls per month) (Noah Gans G. K., 2003) |

The study required a dataset that has information about every case of a client in the queue: times when the client joined and leaved queue. Turned out, that it is not an easy task to find suitable dataset. The data freely available to use are mostly consists of average time in queue information within fixed intervals of time.

## About call center dataset.

Dataset "Anonymous Bank" Call-Center Data (Israel Institute of Technology, 2002) was used in this work. The data collected by the call center of one of Israel’s banks. This is one of the most documented call-center datasets that employs a lot of detailed data as was mentioned in the article (Lawrence D Brown, 2005). The data contain information about every call that was made to the bank for all months in 1999. Each entering phone call is first routed through a VRU (Voice Response Unit) where customers identify themselves. Part of the customers (35%) whose problems were not solved (with the help of the VRU) join the queue for operator service, we are interested in this type of call. A detailed scheme of the call center operation can be found in Figure 1.

Description of all data fields can be found in dataset documentation (Israel Institute of Technology, 2002), but not all of them were necessary for this work. Description of fields that were extracted to analysis are following:

* *date* ­— year-month-day.
* *vru\_entry* — time when the call was made.
* *q\_time* — time spent in a queue in seconds.
* *outcome* — this field was used to extract hung-up calls. There are 2 possible outcomes: AGENT — service, HANG — hung up.

## Data preparation

The data is contained in 12 files, one file is one month. The first step was to unite 12 months dataset into one dataset of all year. Then, as usual, the dataset was checked on null data. Since the data is quite old, dates were written in string format, so, the function was written to rewrite all calls vru\_entry field to the datetime format, and a new field vru\_entry\_dt appeared. With the help of pandas method pandas.DataFrame.boxplot (McKinney, 2010), box plot was made to find thresholds. 2000 second was chosen as threshold for field q\_time, so, we work only with calls, the time in a queue of which is less than 2000 seconds.

## Visualizing results

The main practical task is to make plots where time in queue depends on some load factor. In the dataset, we can define 2 load factors — the number of people in the queue and the load on the operator (number of waiting customers per operator).

To solve the task the function was written, its inputs are our dataset and time interval in n minutes, its outputs are lists with values of x-axis and y-axis of plots and plots themself. The function splits data into intervals of n seconds and for every interval count: the number of calls in the interval (that shows how many clients in the queue in the interval); the number of unique operators serves customers; mean, median, and standard deviation of column q\_time; the number of clients per operator in the interval.

Then, load factors (number of people in queue and number of customers per operator) add to lists of values of the x-axis and factors of time in the queue to lists of values of the y-axis. This data visualizing to scatterplots with the help of function matplotlib.pyplot.scatter (Hunter, 2007). To find the best graph for interpretation, three intervals — 15, 30, and 60 minutes — were used as inputs to the function. 60 minutes interval was chosen as best and the plots with this interval are presented in the work.

To find the moment from what the chaos begins, plots were added 25th, 50th, and 70th percentiles. The difference between 25th and 75th percentiles shows us how widely scattered are the values. Percentiles calculate the following way. For every unique x\_value of load factor, were considered values of Y which corresponde interval of X: [x\_value - step, x\_value + step]. The step is chosen where about 400-600 dots are in interval. These values of Y are ordered, and percentile ranks counted. Then, lines of percentiles are added to the plot.

An example of the resulting graph is shown in Figure 2. The blue dots here are one-hour intervals, their value X is the number of people in the queue per hour, and Y is the average waiting time in the queue during this hour. The lines on the plot show how percentiles change for the nearest 100-200 hour intervals on the chart. To make plots neater, there were thresholds for load factor: maximum 20 people per operator and maximum 200 customers in the queue. Also, there were set the y-axis view limit which corresponds to (maximum of 75th percentile).

|  |
| --- |
| Map  Description automatically generated with medium confidence |
| Figure 2 The example plot |

# Results

## Waiting time metrics

The load plots are shown in Figure 3. Here, the load factor is the number of people in the queue and the number of customers per operator. For these load factors, the mean, median, and standard deviation were calculated.

The increasing difference between 25th and 75th percentiles suggest the system is no longer able to handle the load. Denote, that waiting time in the queue also increase with increasing load, but it is the growth of differences between percentiles that shows us that chaos began in the system. Some calls are served fairly quickly, while some others are terribly slow.

Let’s pay attention to plots of mean and median waiting time. It can be seen that growth slows down starting from 15 people per operator and 150 people in the queue. Suppose that the reason for the slowdown in growth is that people wait too long that they hung up. In (Lawrence D Brown, 2005) were shown that there is a strong linear relationship between average waiting time and abandoned calls in the queue fraction. We checked this assumption by making plots of the average waiting time of people before hung up depends on operator load and the number of abandoned calls depends on operator load. Zero values of average waiting time were deleted. The resulting plots are in Figure 4. Our guess was confirmed, when operators load becomes 15 people per operator or more, the number of hung ups greatly increases by nonlinear dependence.

|  |
| --- |
| Diagram  Description automatically generated |
| Figure 3 Plots of waiting time with different load factors |

Looking at the plot of average waiting time of callers before hanging up, we see that at the beginning values of y-axis growths, but then becomes about the same with increasing load. I think this is because at the beginning the system can handle the load, so, most people don't have to wait long, but there is a time limit people ready to wait, and when the system reached that limit, the average waiting time stops growing. This remark also may be an explanation of the slowdown in the growth of mean and median waiting time in Figure 3.

## Service quality

Service quality is measured by the following indicators (Noah Gans G. K., 2003):

1. Accessibility of agents. How long does the customer have to wait and how many people have hung up?
2. Effectiveness of service encounter. Did the customer call a second time to ask more questions?
3. Effectiveness of service encounter. How satisfied is the customer with the service?

|  |
| --- |
| Graphical user interface, chart  Description automatically generated |
| Figure 4 Measures of abandoned calls |

The data does not contain information about customers, also, call center customers are not asked to rate the call. Thus, unfortunately, our data allows us to rely only on the indicators of the first point. But even with only known indicators, we can note that in this service system there is a moment when the system can’t handle the load, the moment when operators become less accessible. The system capacity decreases, it becomes unstable, increasing difference between 25th and 75th helps us to reveal this fact. The limit of system capabilities is also confirmed by a big number of abandoned calls and by the limit of waiting time of people before hanging up.

# Testing the Classical Queuing Theory

One of the goals of this paper was to test whether the queuing theory formulas are correct on real data. Call centers are considered to be an M/G/N queue model (**M**arkovian arrivals / **G**enerally distributed service times / **N** servers) where the average waiting time is calculated using the following classical Khintchine-Pollaczek formula (Whitt, 1993):

Here is the coefficient of variation of the service time, N is the number of active operators and denotes the agent’s utilization. The agents’ utilization is also an indicator of the effectiveness of the work and in (Lawrence D Brown, 2005) is calculated using the following formula:

Here is the effective arrival rate (of costumers who get served), is the service rate.

|  |
| --- |
| Chart, scatter chart  Description automatically generated |
| Figure 5 Agents’ utilization vs. average waiting time |

To make the graph shown on the left in Figure 5, first a graph of the average waiting time depending on the utilization for each hour interval was plotted. Then, the intervals along the x-axis were divided into 50 equal parts, the median values were selected for each part. Exactly these medians are represented on the left graph. The dot sizes indicate the number of customers in the interval.

To check whether the Khintchine-Pollaczek formula suits this dataset, the graph shown on the right side of Figure 5 of dependency versus was plotted. According to the formula, the relationship should be linear, but obviously, this is not the case.

This method was taken from (Lawrence D Brown, 2005). The authors got similar results, but it is worth noting that my results are slightly different from theirs, even though they were made on the same dataset. There are already differences that the authors note they got 3,867 hour intervals during the calculations, while this is impossible, since the call center worked 363 days a year, 17 hours a day. The authors also proved that the Khintchine-Pollaczek formula does not fit this dataset. They concluded that this formula is simply not appropriate for queueing systems with abandonment.

# Conclusion

The purpose of our analysis was to study the changes in the efficiency of the queue system with increasing load based on real data. In this work was shown that in this call center there is a moment (approximately at a load of 15 people per operator) when the system proving inadequate and incapable of offering high-quality service. This was reflected in the increasing average waiting time (service delay), the increasing difference of 25th and 75th percentile (service time variability), and the increase in the number of dropped calls with increasing load on the system. All these are indicators of the inefficiency of the call center.

Also, our study that the classical Khintchine-Pollaczek formula is not suitable for this dataset. Perhaps [we should not jump](file:////перевод/английский-русский/we+should+not+jump) to [conclusions](file:////перевод/английский-русский/conclusions) about the inapplicability of the classical formula based on the results of a single dataset, but we can assume that either this formula is not applicable for real queue systems, or for those that have an abandonment.

# References

Israel Institute of Technology. (2002). *Call Center Data.* Retrieved from http://iew3.technion.ac.il/serveng/callcenterdata/index.html

Lawrence D Brown, N. G. (2005). Statistical Analysis of a Telephone Call Center: A Queueing-Science Perspective. *Journal of the American Statistical Association 100(March)*, 36-50.

McKinney, W. o. (2010). Data Structures for Statistical Computing in Python. *9th Python in Science Conference*, (pp. 51-56). Austin, TX.

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in science \& engineering, 9(3)*, 90-95.

Noah Gans, G. K. (2003). Telephone Call Centers: Tutorial, Review, and Research Prospects. *Manufacturing & Service Operations Management 5 (2)*, 79-141.

Lawrence D. Brown, N. G. (2005). Statistical Analysis of a Telephone Call Center. *Journal of the American Statistical Association, 100 (469)*, 36-50.

Noah Gans, N. L. (2010). Service times in call centers: Agent heterogeneity and learning with some operational consequences. *Institute of Mathematical Statistics Collections, 6*, 99-123.

Mohammad Delasay, A. I. (2016). Modeling Load and Overwork Effects in Queueing Systems with Adaptive Service Rates. *Operations Research , 64*(4), 867-885.

Robbins, T. &. (2006). Evaluating Arrival Rate Uncertainty in Call Centers. *Winter Simulation Conference*, (pp. 2180-2187).

Saltzman, R. &. (2005). A manager-friendly platform for simulation modeling and analysis of call center queueing systems. *Simulation Conference*, (pp. 466-473).

Pichitlamken, J. &. (2004). Modelling And Simulation Of A Telephone Call Center. *Simulation Conference*, (pp. 1805 - 1812).

Kendall, D. G. (1953, September). Stochastic Processes Occurring in the Theory of Queues and their Analysis by the Method of the Imbedded Markov Chain. *Ann. Math. Statist., 4*, 338-354.

Erlang, A. (1909). The Theory of Probabilities and Telephone Conversations. *Nyt Tidsskrift for Matematik B, 20*, 33.

Erlang, A. (1920). Telephone waiting times. *Matematisk Tidsskrift B, 31*, 25.

Ludwig Kuntz, R. M. (2011). Stress on the Ward – An Empirical Study of the Nonlinear Relationship between Organizational Workload and Service Quality. *Ruhr Economic Papers, 277*.

Mohammad Delasay, A. I. (2014). The Influence of Load on Service Times.

Whitt, W. (1993). Approximations for the GI/G/m queue. *Production and Operations Management, 2*(2), 114-161.