Statistical Analysis of Call Center data

RESEARCH PROPOSAL

Khudyakov Polina

Professor Paul D. Feigin and Doctor Malka Gorfine

Faculty of Industrial Engineering and Management
Technion - Israel Institute of Technology

October 2007
Contents

List of Acronyms ................................................. 4

1 Introduction .................................................. 7

2 Literature Review ........................................... 9
   2.1 Descriptive Statistics ................................. 9
   2.2 Customer Patience ....................................... 10
      2.2.1 Survival Analysis ................................. 10
      2.2.2 Frailty Models ..................................... 11
      2.2.3 Copulas ............................................. 12
      2.2.4 Bayesian Analysis of Survival Data .......... 12
      2.2.5 Markov and Semimarkov Models ............... 13
   2.3 Retrial Queues ........................................ 13
      2.3.1 Redials after Encountering a Busy Signal .... 13
      2.3.2 Redials after Abandoning ....................... 14
      2.3.3 Revisits after Service ........................... 15

3 Preliminary Analysis ....................................... 17
   3.1 Description of the Data .............................. 17
   3.2 Data Collecting and Processing .................... 19
   3.3 Customer Patience ..................................... 20
   3.4 Customer Retrials .................................... 23

4 Customer Patience Analysis ............................. 26
   4.1 Frailty Models ......................................... 26
      4.1.1 Shared Frailty Model ............................ 26
      4.1.2 Two-level frailty Models ....................... 28
      4.1.3 Frailty as a Stochastic Process ............... 30
   4.2 First-passage-time models .......................... 32
4.3 Copulas ................................................................. 33
4.4 Bayesian Survival Analysis ................................. 34

5 Retrial Analysis .......................................................... 37
  5.1 Customer Retrial Behavior ................................. 37
    5.1.1 Possible Call Scenarios ............................... 38
    5.1.2 Retrial Definition ..................................... 39
  5.2 Multiserver Queues ........................................... 42
    5.2.1 The Open Jackson Network ......................... 42
    5.2.2 Time Varying Multiserver Queues ................. 44
    5.2.3 Finite State Markov Chain ......................... 46
    5.2.4 The Closed Jackson Network ....................... 48
    5.2.5 Comprehensive Model ............................... 49
  5.3 Additional Retrial Analysis .............................. 50

6 Research Goals ........................................................ 51
List of Acronyms

IVR Interactive Voice Response
ACD Automatic Call Distributor
FCFS First Come First Served
LT Laplace transform
IEEE Institute of Electrical and Electronics Engineers
MCMC Markov chain Monte Carlo
i.i.d. independent identically distributed
List of Tables

3.1 The distribution of the number of retrials in a “series of retrials”. 25

5.1 Transition probabilities of moving from the current outcome at a
given attempt to the next state. 41
List of Figures

3.1 Schematic diagram of a customer call ....................... 18
3.2 Smooth hazard function for customer patience while waiting for “Platinum” service provided by company “C” on weekdays in July 2004 ......................................................... 21
3.3 The distribution of customer patience in case of abandonment as calculated for complete “series” ................................................................. 22
3.4 The distribution of customer patience in case of abandonment as calculated for incomplete “series” ................................................................. 22
3.5 Distribution of time (in days) between the first and the last calls in a “series” ................................................................. 23
3.6 Distribution of time (in hours) between the first and the last calls in a “series” ................................................................. 24
3.7 Distribution of time (in minutes) between the first and the last calls in a “series” ................................................................. 24
5.1 Markov chain for the description of a customer call ............ 39
5.2 Customer retrial behavior scheme .................................. 40
5.3 Markov chain for a queue with abandonments, retrials and revisits ................................................................. 43
5.4 The queue with abandonments, retrials and revisits .......... 43
5.5 Markov chain for a queue with abandonments, retrials and revisits ................................................................. 43
5.6 Closed Jackson network for the queue with abandonments, retrials and revisits ................................................................. 48
5.7 Open Jackson network for the queue with abandonments, retrials and revisits ................................................................. 49
Chapter 1

Introduction

In our increasingly industrialized and globalized world, a large number of companies include call centers in their structures. Call centers are intended to provide and improve customer service, marketing, technical support, etc. For a customer, addressing the call center actually means addressing the company itself, and any negative experience on the part of the customer can lead to the rejection of company products and services.

Hence, it is very important to ensure that a call center functions effectively and provides high quality service to its customers. Call centers collect a huge amount of data, and this provides a great opportunity for companies to use this information for the analysis of customer needs, desires, and intentions. Such data analysis can help improve the quality of customer service and lower the costs.

As a part of on-going study, we deal with the analysis of customer patience and customer retrials. Let us consider each of these parameters in turn.

We define customer *patience* as the ability to endure waiting for service. This human trait plays an important role in the call center mechanism. As mentioned above, every call can be considered as a possibility to keep or lose a customer, and the outcome depends on the customer’s satisfaction. Moreover, customers are likely to remember one disappointing service experience better than 20 good ones. From this point of view, an abandoned call is a negative experience which affects the future customer’s choice.

By *retrials* we understand the customer’s repeated attempts to receive the desired service after the initial failure to get it. The failure can be due to the lack of available resources or to unsatisfactory service. In the literature, such calls are also referred to as repeated calls or re-attempts. The existence of retrials after a service terminated call could indicate customer dissatisfaction with the service
received. The negative impact of customer retrials is the increase in system load and, hence, the deterioration of system performance and the corresponding increase in expenses.

In our study, we combine the analyses of patience and retrials because these aspects are highly correlated. Impatient customers often abandon the system, the abandonment can lead to retrials, and, on the other hand, retrials and previous call experience can affect customer patience.
Chapter 2

Literature Review

2.1 Descriptive Statistics

Papers [46], [20], [34] and [32] written in the 70-s are dedicated mostly to the description and analysis of models with customer abandonments and retrials that took place as a result of telephone network impairments. The underlying research was initiated by companies providing telephone services and telephone equipment.

In [46], Roberts analyzed the experimental data of subscriber retrials and the patience needed in making these repeated attempts in order to analyze the post-dialing delay. Collecting the data for this research was time-consuming and required the use of sophisticated equipment. For the analysis of the retrial phenomenon, the author employed the perseverance function, $H(x)$, that gives the proportion of calls failing at the $x$-th attempt and still continuing to $(x+1)$-th attempt. Calculations of this function for all call types showed that the subscriber’s perseverance strongly depends on the cause of failure, but also systematically increases with the attempt number for a given cause of failure.

In [20], Duffy and Mercer showed that the cumulative influence of network performance and customer behavior tends to mould customer attitudes about the telephone network. It also plays an important role in determining the amount of the equipment needed to serve all customer requests and the company revenues from successfully completed calls. The research goal was to identify and solve network problems in order to design properly the network and to meet specified performance criteria. The authors demonstrated that a subscriber type affects his/her behavior on accomplishing the retrial sequence. Business and residence traffic is defined by the day of the week and the time of the day when the call
was made, as well as by the calling distance. It was also shown that the customer holding time after the onset of the call and an indication of a failure have a significant effect on the total time spent on retrials and, hence, on the network load.

The study [34] by Liu can be considered a continuation of the survey conducted in [20]. Liu’s main goal was to provide a comprehensive characterization of network performance and customer behavior in setting up a customer’s desired telephone connection. For these purposes, the author collected data from one of the switching entities in the Bell System network. Based on these data, he summarized different statistical characteristic, i.e. initial attempt disposition probabilities, retrial probabilities, the number of additional attempts, ultimate success probabilities and distribution functions for retrial intervals following different types of incomplete initial attempts. The results obtained in this study can be applied to many network planning applications.

Kort [32], described models and methods developed at Bell Laboratories to evaluate customer acceptance of telephone connections in the Bell System Public Switched Telephone Network. The author presented customer opinion models for transmissions and call setup impairments, customer abandonment and retrial models as well as for a complaint rate model. These models were created to provide a basis for IEEE standards for telephone network performance specifications in a multi-vendor environment. In addition, Monte Carlo simulation methods were used to integrate customer opinion behavior models with network performance characterization models as well as to determine overall opinion distributions and the incidence of abandonment, retrial and complaints.

\section{2.2 Customer Patience}

\subsection{2.2.1 Survival Analysis}

The first model for customer patience was constructed by Palm in 1943 (see [44]). He introduced a so-called time-dependent inconvenience function that is actually a proportional hazard rate function. An important result, postulated by Palm, is the presence of correlation between a hazard rate of the customer patience time and his/her irritation caused by waiting. Palm also suggested that patience was characterized by a Weibull distribution, a specific case of this distribution being an exponential distribution widely used in queuing theory (Erlang-A queue
Hougaard [28] described models of survival analysis with a reference to the body of technical literature available in the field. His book covers the following fields: standard methods for univariate and multivariate survival analysis, multi-state models, shared frailty models, the analysis of recurrent events and competing risk models.

### 2.2.2 Frailty Models

Aalen [1] presented a description of frailty models and distinguished between three types within these models: a model with frailty as a random variable, a model with a frailty built in over time, and a stochastic process model with the event in question corresponding to absorption in a particular state.

Grambsch and Therneau [25] wrote a hand-book on the frailty approach. They presented a variety of examples analyzed with the help of SAS macros and S-Plus functions.

Aalen and Gjessing ([2] and [4]) considered extended frailty models and first-passage-time models. In [2], the authors investigated the shape of various hazard rates. They mentioned the tendency in the literature to study monotone hazard rates, but maintained that, in practice, the shapes of hazard rates often did not follow those simple patterns. Further, they explained the non-monotone phenomenon by modeling survival distributions as the first-passage-time process with the help of a discrete-space Markov chain (Phase Type models), continuous-space diffusions and Wiener processes. In [4], Aalen et al argued that the theory of stochastic processes should be used in event history analysis. They considered some specific examples: Markov chains, martingale-based counting processes, birth type processes, diffusion processes and Lévy processes.

Aalen and Moger [5] presented a hierarchical frailty model constructed by randomizing scale parameters, corresponding to Levy processes, in the Levy frailty distributions. The considered model allowed individual variation due to unobserved individual covariates, and variation due to unobserved common covariates. The authors illustrated the model using the data on post-perinatal infant mortality in siblings from the Medical Birth Registry of Norway.
2.2.3 Copulas

A brief description of an approach based on copulas, with bibliographic comments, was presented by Hougaard in [28].

Quesada-Molina, Rodrígues-Lallena and Úbeda-Flores in [45] considered the main results of the theory, various examples of copula families, and some open problems.

Nelsen [41] presented a detailed introduction to copula theory and concepts, methods of copula construction and special cases of these models.

Shih and Lous [49] investigated two-stage parametric and two-stage semi-parametric estimation procedures for the association parameter in copula models for bivariate survival data where censoring in either or both components is allowed. The authors derived asymptotic properties of the estimators and compared their performance by simulation. Finally, the methods proposed in the paper are applied to an AIDS data set for illustration.

Bandeen-Roche and Chen in [12] proposed exploratory methods for diagnosing the appropriateness of an underlying copula model for bivariate failure time data, allowing censoring in either or both failure times. The authors found that the proposed approach effectively distinguished gamma from positive stable copula models when the sample is moderately large or the association is strong. Data from the Women’s Health and Aging Study were analyzed to demonstrate the proposed diagnostic methodology.

Schmitz [47] addressed the dependence structure of the minimum and maximum of \( n \) i.i.d. random variables \( X_1, \ldots, X_n \) by determining their copula, and proofed their asymptotic independence. Kendall’s \( \tau \) and Spearman’s \( \rho \) for \((X_{(1)}, X_{(n)})\) were also calculated.

2.2.4 Bayesian Analysis of Survival Data

Chen, Ibrahim and Sinha [15] presented a comprehensive treatment of Bayesian survival analysis. The authors discussed state-of-the-art methods for fitting Bayesian survival models. In their work, theory is well balanced with applications thanks to detailed examples and analyses for each of the considered models.

Arbeev, Locatelli, Wienke and Yashin [8] examined correlated frailty models, especially the behavior of the parameter estimates when using different estimation strategies. The authors considered three different frailty models: the gamma model and two versions of the log-normal model, and compared the results of traditional maximum likelihood procedure of parameter estimation with maximum
likelihood methods based on numerical integration and a Bayesian approach using MCMC methods.

2.2.5 Markov and Semimarkov Models

Customer patience can be modeled in a way analogous to that of creating any lifetime model, and, specifically, to a device lifetime model. This makes Cinlar’s [16] paper particularly interesting to us. Their article is dedicated to choosing an appropriate model for the device deterioration process. The authors discussed Markov and semimarkov models of deterioration and showed that general models could be obtained with the help of a standard Poisson random measure and an appropriative choice of several deterministic functions. They also computed the distribution of the device lifetime under a threshold mechanism for failure.

2.3 Retrial Queues

A detailed overview of models that had been investigated before 1997 was made by Falin and Templeton in [21]. A classified bibliography of research on retrial queues in 90-s was made by Artalejo in [9]. In this section, we will refer to some publications that relate to the models under consideration.

In [22] Gans, Koole and Mandelbaum provided a survey of literature on queuing models for call center design. According to this paper, it is possible to distinguish three modes of returning to the call center: after a busy signal, after abandonment and after a service. In our study, we are going to use the classification suggested by these authors.

2.3.1 Redials after Encountering a Busy Signal

Most of the papers dedicated to the theory of retrial queues deal with redials after encountering a busy signal. However, resources of modern call centers allow for extending the number of trunk lines according to customers’ demand. Retrial models with a finite capacity are more relevant for companies specializing in communication networks. For instance, such a model was considered by Tran-Gia and Mandjes in [39] where they suggested a performance analysis of cellular mobile networks. These authors considered a Markov chain model for retrials in a system with a finite customer population. They found an efficient recursive solution for this model and presented a guard channel concept for customer priorities traffic.
2.3.2 Redials after Abandoning

A Markovian multiserver queuing model with time dependent parameters was considered by Mandelbaum, Massey, Reiman, Rider and Stolyar in [37], [36] and [38]. This model is based on the assumption that customers waiting for service may abandon their call and retry later. Using results from [35], the authors provided fluid and diffusion approximations for both the queue length and the virtual waiting time processes. In addition, the obtained approximations were compared to simulations, and their performance proved to be extremely good.

The phenomenon of retrials in a call center with multiple servers was also analyzed by Aguir, Karaesmen, Aksin and Chauvet in [6]. These authors extended the previous model to include customer balking behavior and finite system capacity. The model analysis was accomplished using a continuous Markov chain and a fluid approximation. The authors analyzed the system for both single- and multi-period settings. For the former case, a fluid approximation was proposed to estimate the retrial rate. For the latter, a numerical analysis was required because the system in this case was intractable. The authors also compared these two overall systems and showed that there was a significant impact of arrival rates and staffing distribution on the retrial rate.

Analysis of time-dependent contact center models with retrials was also provided by Henken in her Ph.D. thesis (see [27]). The author used a fluid approach and diffusion refinement in order to show the influence of retrials on a call center performance and on profitability measures for different routing scenarios.

A call center analysis with recommendations of how to manage the center was presented by Harris, Hoffman and Saunders in [26]. In this paper, the authors analyzed the toll-free nationwide telephone system of the International Revenue Service (IRS). By using simulations and a sensitivity analysis, the authors provided a multiserver loss/delay queue with retrials and reneging. Harris et al developed an effective way of providing reasonable service quality at minimum costs. The resulting model allowed IRS to determine the optimal configuration of the staff and the number of telephone lines for any expected level of incoming traffic.

There are several possibilities to model a call center with retrials. The choice of an appropriate model depends on the problem to be solved and the possibility of finding a solution. Generally, most convenient models for such an analysis are of open type, i.e. they do not have restrictions on the number of places in the
system. Such models were considered previously in [38], [6] and [26]. However, in some cases, it is reasonable to use a closed model, i.e. a model with a limited number of places. For instance, Jennings and de Vericourt in [30] dealt with the problem of hospital staffing when they had to take into consideration the number of places in the system, namely, an always finite number of beds in a given hospital. Another type of closed model was considered by Randhawa and Kumar in [33]. Their system was limited to a number of subscribers. As mentioned above, such a model is appropriate for communication systems.

### 2.3.3 Revisits after Service

De Vericourt and Zhou [50] considered retrials generated by service quality related factors. They analyzed the routing problem in the system, in which customers call back when their problems are not completely resolved by the customer service representatives. The authors integrated the service quality related information into call routing decisions. This procedure allowed one to minimize the average total time for call resolution, which is defined by the authors as the total customer time spent in the system in order to resolve one issue and which includes all related callbacks. This measure of call quality was called resolution probability. The authors did not deliberate on the calculation method and only mentioned that the probability could be defined and measured on the basis of numerous data accumulated by modern call center systems.

As previously discussed in Section 1, the retrial phenomenon increases the system’s load and brings about an additional effort, ultimately increasing expenses. However, the retrial phenomenon is not always negative. Thus, Bucklin and Sismeiro [14] developed and estimated a model for the browsing behavior of a Web site visitor. From the point of view of Web site owners, the existence of retrials is a desired state because it means that customers do not lose their interest in the site. Bucklin et al examined two basic aspects of customer behavior within a site system:

1) making a decision to continue browsing (by submitting an additional page request) or to exit the site;

2) inclination to spend some time viewing each page.

In terms of a retrial model, the first aspect reflects the possibility that a customer will return to the system later, while the second relates to the time that passes between retrials.
The authors proposed a tobit model of type II that captured both aspects of browsing behavior and handled the limitations of the server log-file data. Empirical results showed that customer propensity to continue browsing changed dynamically as a function of the depth of given site visits and the number of repeated visits to the site. It was found that repeated visits led to reduced page-view propensities, but not to reduced page-view durations. The results also revealed browsing patterns that could reflect customer time-saving strategies. The authors also reported that simple site metrics computed at the aggregate level diverged substantially from individual-level modeling results.
Chapter 3

Preliminary Analysis

3.1 Description of the Data

The data we analyze are provided by two call centers belonging to communication and financial companies, which we will address as company “C” and company “F”, respectively. For call center “C” the data have been collected over a period of more than 3 years (April 2004-present), and for call center “F”, we have the data covering a 4 month period (October 2006-January 2007). Both call centers work 24 hours a day on weekdays (Sunday-Thursday). They close at 13:00 on Friday, and reopen at about 17:00 on Saturday. A customer making a call to either call center, receives the service through an IVR or directly from an agent. After receiving the service provided by an IVR, the customer leaves the system or requests service from an agent. The customer requesting service from an agent (after IVR service or from the outset) is redirected to the agent pool. If all agents are busy, the customer waits in a queue. Otherwise, he/she is served immediately. The customer is not always ready to wait in a queue, and he/she can choose to abandon the system at any point during the waiting period. After being served by an agent, the customer either finishes the call or proceeds to another service (another agent), and so on.

\footnote{We do not disclose real names of the companies for confidentiality reasons}

\footnote{In the near future, we will receive the current data for the period beginning from February 2007}
Figure 3.1 presents a schematic diagram of a customer call. This diagram shows that each call starts at an IVR node, and the transition from one agent service to another agent service happens via the IVR node. In practice, the transfer from one agent to another does not always require an IVR service. For such cases, we consider an IVR service time as equal to zero. The service received through an IVR or an agent is reported by the ACD in a separate row named “call segment”. The waiting time is always contained in the same row as the requested service. All calls consist of a number of “call segments”. For example, the first segment describes time in an IVR, the second describes the service, etc.

The considered call centers provide different kinds of services. Some of them are very similar in design and in the average service time. Others, on the contrary, are conceptually different. In our analysis, we will combine services of a similar kind into one group.

The data does not contain any personal information about customers, such as age, social status, family status, education etc. This way, our analysis will be done only on the basis of technical characteristics of the call. For each call, we have the following data:

- the information about a customer initiating this call (customer identification number and customer priority),
- the beginning of each “call segment”,

...
• the duration of each stage of a “call segment” (the service time, the waiting
time in a queue or the post-call agent service time),
• the type of the service (an IVR service or an agent service that can include
about 15 different subtypes),
• the classification of call termination (after a received service, after call aban-
donment or due a system error).

3.2 Data Collecting and Processing

We identify each customer by his/her identification number which is retained in
the field named “customer_id”. It is important to mention that in the data of
company “F”, the field “customer_id” provides a unique number, while in the
data of company “C”, it reflects the telephone number, from which the call was
made. It is only natural to suppose that the identification of the customer by
telephone number is not really reliable because the customer can make a call
from any phone, while on the other hand, a number of customers can use the
same phone. In order to avoid this inconsistency, we analyze only calls made
from cellular phones, and this restriction can be justified on the assumption that
customers prefer to make calls from the phone they most often have at hand.
Moreover, in the data for both companies, “customer_id” can be unidentified or
invalid. To avoid fake I.D. numbers we consider the data only for customers with
fewer than 30 calls a month.

We suppose that customers who abandoned their calls within 2 seconds do
not really intend to receive the service. Thus, we ignore all calls with waiting
time less than 3 seconds.

We provide analyses of two types:

1) the analysis of abandonment and customer behavior following the abandon-
ment

2) the analysis of retrials.

In the analysis of type 1, we use a notion of a “series” which we define as a
sequence of one customer calls happening in chronological order. Each “series”
begins with the first abandoned call and ends with the nearest non-abandoned
call. We distinguish between complete and incomplete “series”. The former is a
“series” with the service received at the last attempt, and the latter is a “series” with the last attempt that also failed.

The results of the analysis of type 2 are used later, in Chapter 5, to build models. In his analysis, we introduce a new term, namely, a “series of retrials” and arrange the data in the following way: For a one-month period, we consider all calls of company “C” customers who were identified by the system. In order to identify a retrial, we make the following assumptions:

- If a customer requested an agent’s service during an initial call, and the time-interval until the next call is less than 2 days then the second call is a retrial.
- If a customer did not request the agent’s service during the initial call, we consider the retrial interval to be less than 1 day.

This way we obtain the sequences of calls occurring in chronological order and having different call outcomes. We suppose that the calls within the same “series of retrials” are made by the customer in order to complete one specific task. This task can include different issues, but they all relate to the same query.

### 3.3 Customer Patience

The assessment of customer patience is a complicated issue because in most cases customers receive the required service before they lose their patience. The data with non-zero service time are called censored data, and these data require analysis of a special kind, known as survival analysis. For the purpose of preliminary analysis, we illustrate the changes in customer patience with the help of the hazard function. Adjusting the definition in [28] to our case, at any time point $t$, the hazard function is defined as the probability of abandonment within a short interval, given that the customer was in a queue at the beginning of the interval.
Figure 3.2: Smooth hazard function for customer patience while waiting for “Platinum” service provided by company “C” on weekdays in July 2004.

Figure 3.2 presents smooth hazard function for customer patience while waiting for “Platinum” service provided by company “C” on weekdays in July 2004. It is actually the Kaplan-Meier estimate of the hazard function smoothed by the super-smooth algorithm. The behavior of the function shows that customer patience distribution is not a monotone function, and it can be considered as a process developing over time. Methods for more profound survival analysis of customer patience are discussed in details, in Chapter 4.

Now let us consider customer patience in the case of abandonment with the help of a number of call “series” constructed in the way described in the previous section. For this purpose, we use company “F” data for October 2006. Let us assume that the maximum time of the waiting interval between the calls in the “series” is the customer patience in case of abandonment. Its distribution was calculated for complete (49%) and incomplete (51%) “series”.

21
Figure 3.3: The distribution of customer patience in case of abandonment as calculated for complete “series”.

Figure 3.4: The distribution of customer patience in case of abandonment as calculated for incomplete “series”.

Figures 3.3-3.4 present distributions of customer patience in case of abandonment for complete and incomplete “series”. In general, the conditional distributions presented above are similar: about 25% of the customers abandoned the system within the first 30 seconds, after that there was a small peak between 1
and 1.5 minutes, and then the distributions start to decrease. However, this peak is less pronounced for the incomplete “series” than for the complete one. There is also another subtle difference; the percent of customers who abandoned within the first 30 seconds for the complete “series” is about 22%, while for the incomplete “series” it is 28%. This difference can be explained by the assumption that customers less interested in the receiving the service will be ready to abandon the system within a shorter period of time. Moreover, they rarely redial to get the service.

3.4 Customer Retrials

We start this section by presenting the distribution of time between the first abandonment on the part of the customer and the service received in some time after this abandonment. In terms of our study, this distribution means the time distribution between the first and the last calls in a complete “series”. As before, we use company “F” data for October 2006.

![Distribution of time between the first and the last calls in a "series".](image)

Figure 3.5: Distribution of time (in days) between the first and the last calls in a “series”.

Figure 3.5 shows that about 40% of customers are served during the same day, 10% receive their service on the next day. The residual of 10% receive service beyond the 10-day period of time, and we can assume that the received service
is not related to the one requested in the previous call.

Figure 3.6: Distribution of time (in hours) between the first and the last calls in a “series”.

Figure 3.6 shows that customers served within a day of the primary call were actually served within the first hour.

Figure 3.7: Distribution of time (in minutes) between the first and the last calls in a “series”.

24
Figure 3.7 shows that customers who redial within the first hour after the primary call are actually served at their second call, made straight after the first one.

As mention previously in section 3.2, the second type of preliminary analysis is intended for construction of models in chapter 5. Here, we only present the table that demonstrates the existence of retrials in our context.

<table>
<thead>
<tr>
<th>Number of attempts</th>
<th>Number of customers</th>
<th>% of customers</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>199661</td>
<td>57.04</td>
</tr>
<tr>
<td>2</td>
<td>74178</td>
<td>21.19</td>
</tr>
<tr>
<td>3</td>
<td>33538</td>
<td>9.58</td>
</tr>
<tr>
<td>4</td>
<td>17137</td>
<td>4.90</td>
</tr>
<tr>
<td>5</td>
<td>9488</td>
<td>2.71</td>
</tr>
<tr>
<td>6</td>
<td>5567</td>
<td>1.59</td>
</tr>
<tr>
<td>7</td>
<td>3425</td>
<td>0.98</td>
</tr>
<tr>
<td>8</td>
<td>2120</td>
<td>0.61</td>
</tr>
<tr>
<td>9</td>
<td>1446</td>
<td>0.41</td>
</tr>
<tr>
<td>More</td>
<td>3501</td>
<td>1.00</td>
</tr>
<tr>
<td><strong>Total:</strong></td>
<td><strong>350061</strong></td>
<td><strong>100</strong></td>
</tr>
</tbody>
</table>

Table 3.1: The distribution of the number of retrials in a “series of retrials”.

Table 3.1 shows that about 60% of customers do not call back within the first two days, about 20% of customers make one retrial 10% make two retrials. This finding proves the existence of retrials, and gives us a motivation for the analysis.
Customer Patience Analysis

This chapter provides a review of various statistical models that can be used in analysis of the customer patience, under the setting of survival analysis. In terms of our context, an \textit{event} is the customer abandonment of the system before being served. Customer who gets served, namely his/her patience time is not fully observed is considered as \textit{censored}. Hence, for each customer, at each call, the observed time is the time until abandonment (patience time) or time until being served, whichever comes first. Parts of this chapter are based on Hougaard [28].

4.1 Frailty Models

The data to be used in the current research consists of customer calls with possibly multiple calls for a customer (see 3.2 for details). Hence, it is reasonable to assume dependence between the observed times of the same customer. In this situation, the Cox proportional hazard model [17] cannot be used directly, and a well-known approach that deals with clustered data is by using frailty models.

4.1.1 Shared Frailty Model

In our context the assumption of \textit{shared frailty model} is that calls of the same customer (cluster) share the same frailty variate \(Y\). The customer patience times (\textit{survival times}) are assumed to be conditionally independent with respect to the shared frailty. Conditional on the unobservable frailty variate \(Y\), the hazard function of customer patience at the \(j\)-th call is assumed to take of the form

\[
h_j(t|Y) = h_0(t)Ye^{\beta z_j}, \quad j = 1, ..., m_i, \tag{4.1}
\]
where $h_0(t)$ is an unspecified baseline hazard function, $z_j$ is the vector of observed covariates, $\beta$ is a vector of unknown regression coefficients and $m_i$ is the number of calls of customer $i$, $i = 1, \ldots, n$, where $n$ is the number of customers. We assume that $Y_i$, $i = 1, \ldots, n$ are i.i.d. random variables with density $f(y) \equiv f(y; \theta)$, where $\theta$ is an unknown vector of parameters. On the other hand, $Y_i$ is common for all calls of the same customer, and thus induces dependence between survival times within calls of a customer. This model is also called a common risks model. For simplicity of presentation, we omit the index $i$ wherever there is no confusion.

The conditional bivariate survival function conditioning on $Y$ is expressed as follows

$$S(t_1, t_2|Y) = \exp\{-Y (M_1(t_1) + M_2(t_2))\}, \quad (4.2)$$

where $M_j(t) = \int_0^t h_0(u)e^{\beta^T z_j}du$, $j = 1, 2$ are the integrated conditional hazards. From this, the unconditional bivariate survival function is immediately derived by integrating $Y$ out, so that

$$S(t_1, t_2) = E\left(\exp[-Y (M_1(t_1) + M_2(t_2))]\right) = LT\{M_1(t_1) + M_2(t_2)\}, \quad (4.3)$$

where $LT(s)$ is the Laplace transform of the distribution of $Y$. Letting $M(t_1, \ldots, t_{m_i}) = \sum_{j=1}^{m_i} M_j(t_j)$, the multivariate survival function of $m_i$ observations is

$$S(t_1, \ldots, t_{m_i}) = LT\{M(t_1, \ldots, t_{m_i})\}. \quad (4.4)$$

As an alternative, all formulas can be expressed by means of the marginal distributions $S_j(t) = P(T_j > t)$. The bivariate unconditional survival function corresponding to (4.3) is

$$S(t_1, t_2) = LT\left[LT^{-1}\{S_1(t_1)\} + LT^{-1}\{S_2(t_2)\}\right], \quad (4.5)$$

and its multivariate version is

$$S(t_1, \ldots, t_{m_i}) = LT\left[\sum_j LT^{-1}\{S_j(t_j)\}\right] \quad (4.6)$$

since $LT^{-1}\{S_j(t_j)\} = M_j(t_j)$.

One important problem in applying frailty models is the choice of the frailty distribution. The widely used distributions are gamma, positive stable, log normal, log t, inverse Gaussian and compound Poisson. Each distribution family has different properties, thus, for example, the stable frailty distribution implies high early dependence, whereas the gamma frailty model describes high late...
dependence. A detail comparison of the properties of different frailty families can be found in [28]. For fitting a model, the parameters $\theta$, $\beta$ and $H_0(t|Y) = \int_0^t h_0(u|Y)du$ have to be estimated. There are different estimation methods such as EM-algorithm based approach, a two-stage approach, a penalized likelihood approach, among others. Each estimation procedure has its own pros and cons. A detailed description for the above mentioned methods and their comparison are presented in [28] and references therein. In this study we intend to use different frailty distributions, and compare their fit to our data.

As mentioned in section 3.1, we can use only technical call details as covariates for customer patience. Most probably, they could be time of calling (day-time with different call traffic may have various effects on the customer patience), and the type of service requested by the customer (different services may have different service times and are of different importance to the customers).

The shared frailty model (4.1) takes into account the personal factors of a customer, but not his/her experience over the time. However, it is reasonable to assume that the customer calls history influence on his/her current waiting behavior. One of the models dedicated for such analysis is the well studied recurrent events model and its extension to shared frailty model. However, these type of models cannot be applied directly in our case, since in a typical recurrent event data, subject can be censored at most once, and no information is available after this censoring time.

A possible extension of model (4.1) can be the following:

$$h_j(t|Y) = h_{0j}(t)Y e^{\beta_j z_j}, \quad j = 1, \ldots, m$$

where $h_{0j}(t)$ is the unspecified baseline hazard function of call $j$. In this model, the baseline hazard functions are assumed to be different at each call, since it could be that the customer behavior changes as he/she becomes more experienced with the system.

It is also possible to consider a model with different regression coefficient vectors $\beta_j$, with $h_0(t)$ or $h_{0j}(t)$.

### 4.1.2 Two-level frailty Models

A possible disadvantage of model (4.7) is the assumption that customer patience changes with the number of call consistently for all customers, while, it could be that these changes have individual features as well. Thus, we can extend model (4.7) to include two random factors: a frailty variate $Y$ at the customer level, and
a frailty variate $V_j$ at a call level of each customer. Such model was considered by Aalen et al [5]. It assumes that random effects operate multiplicatively on the baseline hazard, and conditional on the frailties $Y$ and $V_j$, the hazard function of customer patience at call $j$ is of the form

$$h_j(t|Y, V_j) = Y V_j h_{0j}(t)e^{\beta^T z_j}, \quad j = 1, \ldots, m_i$$

(4.8)

where, for each customer, $V_j j = 1, \ldots, m_i$ are i.i.d. random variables with density function $g(v) \equiv g(v; \mu)$ where $\mu$ is an unknown vector of parameters. The frailty $V_j$ can also be regarded as the set of covariates of call $j$ that are not included in $z_j$ because they are not measured. This is an hierarchical frailty model with two-levels frailty: shared frailty at the customer level and unshared frailty at the call level of each customer.

In our preliminary analysis in Chapter 3 we investigate the possibility of grouping the customer’s calls in “series of retrials” (see 3.1). Hence, for studying the effect of “series of retrials” we can consider the following model:

$$h_j(t|Y, W_k) = Y W_k h_{0j}(t)e^{\beta^T z_{kj}}, \quad k = 1, \ldots, l_i, \quad j = 1, \ldots, m_k$$

(4.9)

where $w_k$ is the frailty variate of the $k$-th series of a specific customer, $l_i$ is the total number of series of the customer, and $m_k$ is the total number of calls of series $k$. This is an hierarchical frailty model with two-levels frailty: shared frailty at the customer level and shared frailty at the “series of retrials” level of each customer. This model can be further extended to three-levels frailty model, by adding $V_{kj}$ frailty variate at the level of the call at each “series of retrials” of each customer.

Hierarchial frailty model (4.9) is also considered by Aalen et al [5] under frailty distributions determined by non-negative Lévy processes, which includes PVF distributions. The PVF distributions include the gamma, positive stable, inverse Gaussian and compound Poisson distributions as a special cases. Frailty models specify a parametric dependence structure of the multivariate failure-time data. The parameters’ estimates can be sensitive to the assumption of the assumed model, and hence model fit is an important issue. As mentioned before, Hougaard [28] provides a comprehensive discussion of the theoretical properties and the fit of the PVF distributions under model (4.1). Hsu et al [29] also consider model (4.1) and show by simulation that the biases in the marginal regression estimates and the marginal hazard function are generally 10% or lower under the assumed gamma distribution and mis-specification of the frailty distribution. This suggests that the gamma frailty model can be practical choice if
the marginal parameters are of primary interest. However, when the dependence function is also of interest, a correct specification of the frailty distribution is crucial. A general diagnostic approach to check the bivariate association structure of clustered failure times is given in [24]. Additional tests and graphical procedures for checking the dependence structure of clustered failure-time data can be found in [12], [23], [29], [49] and [51]. These procedures however are not directly applicable to the hierarchical frailty models we deal with here; extension of the procedures to these models will be needed.

Aalen et al [5] proposed parametric choices for the baseline hazard such as Weibull, exponential and Gompertz distributions. The case with nonparametric baseline hazard function can be considered in our ongoing study.

4.1.3 Frailty as a Stochastic Process

The process of waiting on line before being served can be affected by factors developing within the time. For example, at the beginning a customer is expecting to wait a specific period of time, after this period he/she is astonished and after a while even angry for being waiting. In a simple frailty model the frailty variate is constant in time, but may be a more realistic model would be to assume a frailty that developing with time. This can be modeled by considering frailty as a stochastic process. Conditional on the unobserved frailty variate $Y(t)$, the hazard function of the customer patience is of the form

$$h_j \{t|Y(t)\} = h_0(t)Y(t)e^{\beta T z_j}, \quad j = 1, \ldots, m_i. \tag{4.10}$$

Here, $Y(t)$ is a stochastic process.

One of the described models in the literature is the Woodbury-Manton model (see [1] and [28]). It suggested to consider a diffusion process that formulated by a stochastic differential equation:

$$dZ_t = a_0(t)dt + a_1(t)Z_t dt + bdW_t, \tag{4.11}$$

where $a_0(t)$ is a fixed change per time unit, $a_1(t)$ is a change dependent on the value of the process at time $t$ and $W_t$ is a Brownian motion, which is a prototype of random variation. If the frailty at time $t$ is defined as $Y(t) = Z_t^2$, so that the population hazard (the univariate case) rate for this model is given by

$$h(t|Y(t)) = Y(t)h_0(t)e^{\beta T z_j} = \{m(t)^2 + \sigma(t)\}h_0(t)e^{\beta T z_j}, \tag{4.12}$$

30
where \( m(t) \) and \( \sigma(t) \) are the mean and variance of \( Z_t \) conditional on survival until time \( t \) and the solutions of the following differential equations:

\[
m(t)' = a_0(t) + a_1(t)m(t) - 2m(t)\sigma(t)h_0(t)e^{\beta z_j}, \tag{4.13}
\]

\[
\sigma'(t) = 2a_1(t)\sigma(t) + b^2 - 2\sigma^2(t)h_0(t)e^{\beta z_j}, \tag{4.14}
\]

under the initial conditions defined by \( m(0) \) and \( \sigma(0) \). Hence, an estimation technique is required for \( m(t) \), \( \sigma(t) \), \( a_0(t) \), \( a_1(t) \), \( b \), \( \beta \) and \( h_0(t) \). A detailed review of the Woodbury-Manton model and procedures for calculation of marginal survival functions are presented in [52]. Parameters can be estimated from multiple data sets with an assumption that observations are independent given a frailty \( Y \). Our analysis of the literature shows that the model has never been considered for bivariate data, and therefore, it is not known how much more difficult the equations are in that case. The conclusion on this model is that it needs to be developed further and studied in more detail before we can discuss its applicability in our study.

Aalen et al [3] studied models with frailty as a weighted Lévy process. For this purpose, they defined nonnegative deterministic rate function \( r(t) \) with integral \( R(t) = \int_0^t r(u)du \) and let \( Y\{R(t)\} \) to be the time-transformed subordinator. Conditional on \( Y \) the hazard rate process is defined as

\[
h(t|Y) = h_0(t) \int_0^t a(u, t - u)\,dY\{R(u)\}, \tag{4.15}
\]

where \( a(u, t - u) \) is a nonnegative weight function. There are several options to include the covariates in Lévy process model: in the rate function, in the baseline hazard function or in the weight function. More detailed these methods considered in [5].

Special cases of such models are the following:

- Standard frailty model, when \( a(u, t - v) \equiv 1 \) and \( r(t) \) equals \( \rho \) up to time \( T \) and 0 after this time;

- Cumulative frailty model, when \( a(u, t - v) \equiv a_1(t) \) and \( r(t) \equiv h_0(t) \equiv 1; \)

- Moving average frailty model, when \( a(v, t - v) \equiv a_2(v) \) and \( r(t) \equiv h_0(t) \equiv 1. \)

The authors do not discuss how to choose the weight function, but proposed some possible weight functions. According to our knowledge, no work presents estimation technique under this model is available in the literature. Hence, for applying this approach to our data, a first step should be to develop an estimation procedure.
4.2 First-passage-time models

First-passage-time models are stochastic process models where the event in question corresponds to absorption in a particular state. Aalen et al [2] and [4] considered the following first-passage-time models: phase type models, random walk with absorption state and Wiener process with absorption. Consider a Wiener process with random drift. Assume that each individual moves along an axis according to a Brownian motion with drift that is randomized according to a normal distribution with expectation $-\mu$, ($\mu > 0$) and variance $\tau^2$ and let the variance coefficient of the Wiener process equals 1. The process starts at a given point $c$, ($c > 0$) and the absorption state is 0. The time till absorption follows an inverse Gaussian distribution with mixed drift parameter and its density function is the following:

$$f(t) = \frac{c}{\sqrt{2 \pi} t \sqrt{\tau^2 + t}} e^{-\frac{(c-\mu t)^2}{2(\tau^2 + t)}} ,$$

(4.16)

while the survival function (the probability of not being absorbed by time $t$) may be derived as:

$$S(t) = \Phi\left(\frac{c - \mu t}{\sqrt{\tau^2 + t}}\right) - e^{2c\mu + 2\tau^2} \Phi\left(\frac{-c - 2\tau^2 - \mu t}{\sqrt{\tau^2 + t}}\right),$$

(4.17)

where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal distribution. It should be noted that this is a defective survival distribution, because the crossing into the absorbing state may never take place.

This model allows us to distinguish two different types of covariates, namely those which represent measures of how far the underlying process has advanced, and those which represent causal influences on the development. It is natural to model the first type as influencing the distance from absorption that is, the parameter $c$, while the causal covariates influence the drift parameter $\mu$. Thus, one or both of parameters $\mu$ and $c$ are appropriative functions of covariates.

Aalen et al [2] and [4] considered the case of i.i.d. observations with no clusters and suggesting to use maximum likelihood estimates. In our data we have number of measurements for each customer, so, an extension of the model is required. One possible extension is to let the customer drift be the same for all calls, and to assume that given the drift, the customers’ calls are independent. Hence, the contribution of each customer to the likelihood function is a product of factors such that a censored call has a likelihood contribution equal to $S(t)$ and a non-censored call has a likelihood contribution equal to $f(t)$. 

32
Our goal is to develop this model, apply it to our data and compare the results with the models mentioned previously.

4.3 Copulas

Copulas are functions which join one-dimensional marginal distributions into the joined multivariate distribution. A $m$-dimensional copula $C$ is a $m$-dimensional distribution on $[0, 1]^m$ with standard uniform marginal distributions. Sklar’s Theorem (see for example [41]) is the central theorem in the theory of copulas, states that every distribution function $F$ with margins $F_1, ..., F_m$ can be written as

$$F(x_1, x_2, ..., x_m) = C(F_1(x_1), F_2(x_2), ..., F_m(x_m)) \quad (4.18)$$

for some copula $C$, which is uniquely determined on $[0, 1]^m$ for distributions $F$ with absolutely continuous margins. Conversely any copula $C$ may be used to join any collection of univariate distribution functions $F_1, ..., F_m$ using (4.18) to create a multivariate distribution function $F$ with margins $F_1, ..., F_m$.

Copulas became rather popular in financial and actuarial circles, and are used for estimating the probabilities of common defaults or generally to estimate dependent risks. For instance, risk managers use copulas to calculate joint default probability for two or more correlated companies and related securities.

The best known families of copulas are the Frechet family, the Gaussian family and the Archimedian family. As an example, let us consider the Archimedian copula, where the survival function has the following form:

$$S(t_1, t_2) = \psi \left( \psi^{-1}(1 - t_1) + \psi^{-1}(1 - t_2) \right), \quad 0 \leq t_1, t_2 \leq 1, \quad (4.19)$$

where $\psi$ is known as a generator function and any generator function which satisfies the properties below is the basis for a valid copula:

$$0 \leq \psi(1) \leq 1, \quad \psi(0) = 1, \quad \psi'(x) < 0, \quad \psi''(x) > 0.$$

One of the Archimedian family members is the Clayton-Oakes copula in which

$$\psi(x) = (1 - x^\theta)^{-1/\theta}, \quad \theta > 0.$$ 

In this case the copula model is actually a gamma frailty model. In the positive stable frailty case, the survivor function is

$$S(t_1, t_2) = \exp \left( - \left[ \{- \ln(1 - t_1)\}^{1/\alpha} + \{- \ln(1 - t_2)\}^{1/\alpha} \right]^\alpha \right).$$

$^1$Parts of this section are based on [41]
This is the Archimedian copula where function $\psi(s) = \exp(-s^\alpha)$ is the Laplace transform for positive stable distributed value. Thus frailty models where the marginals are assumed fixed, is a particular case of the Archimedian copula (4.19) with $\psi(s) = LT(s)$. Nelsen [43] shows that continuous nonnegative random variables with a Shur-constant joint survival function are characterized by having an Archimedian survival copula, and examine dependence properties and correlation coefficients for random variables with Shur-constant survival functions.

Most of the published works deal with copula models considering the bivariate setting while only a few of them deal with the $m$-dimensional setting. It corresponds with the difficulties in the construction of $m$-copulas (for details see [42]). Another possible disadvantage of a copula model is in the assumption of identical dependence parameter between each pair out of the $m$ random variables. Previously we mention the possibility that customer behavior changes together with his/her call experience, therefore, analogously to Shih et al [49], we can use different marginal distributions for each call $j$ of customer $i$. We also plan to consider the use copula models for the analysis of our data.

4.4 Bayesian Survival Analysis

Bayesian approaches to survival analysis have recently received much attention in the literature due to advances in computational and modeling techniques. Bayesian inference considers parametric and semiparametric models, proportional and nonproportional hazard models, frailty models, models for multivariate survival data and some others.

Let $D$ be the observed data and $\theta$ a vector of unknown parameters. The Bayesian paradigm assumes that $\theta$ is a random variable and has prior distribution denoted by $\pi(\theta)$. The posterior distribution of $\theta$ is given by

$$
\pi(\theta|D) = \frac{L(\theta|D)\pi(\theta)}{\int_{\Theta} L(\theta|D)\pi(\theta) \, d\theta},
$$

(4.20)

where $\Theta$ denotes the parameter space of $\theta$. Therefore, $\pi(\theta)$ is proportional to the likelihood multiplied by the prior,

$$
\pi(\theta) \propto L(\theta|D)\pi(\theta),
$$

(4.21)

and thus it involves a contribution from the observed data through $L(\theta|D)$, and a contribution from prior information quantified through $\pi(\theta)$. The quantity

---

2 Parts of this section are based on [15]
\[ m(D) = \int_{\Theta} L(\theta|D) \pi(\theta) d\theta \] is the normalizing constant of \( \pi(\theta) \), and is often called marginal distribution of the data or the prior predictive distribution. When \( m(D) \) does not have an analytic closed form, usually Markov Chain Monte Carlo (MCMC) sampling algorithms are being used for sampling from \( \pi(\theta) \). One of the most popular MCMC methods is called Gibbs sampler. This is a very powerful simulation algorithm that allows us to sample from \( \pi(\theta) \) without knowing the normalizing constant \( m(D) \). There are novel computational techniques to carry out the posterior computations even if the prior \( \pi(\theta) \) itself does not have a closed analytic form.

A major aspect of the Bayesian paradigm is prediction. The posterior predictive distribution of a future observation vector \( z \) given the data \( D \) is defined as

\[
\pi(z|D) = \int_{\Theta} f(z|\theta) \pi(\theta|D) d\theta,
\]

(4.22)

where \( f(z|\theta) \) denotes the sampling density of \( z \), and \( \pi(\theta|D) \) is the posterior distribution of \( \theta \). It is clear that (4.22) is just the posterior expectation of \( f(z|\theta) \), and thus sampling from (4.22) is easily accomplished via the Gibbs sampler from \( \pi(\theta|D) \). This features of the Bayesian paradigm shows that predictions and predictive distributions are easily computed once samples from \( \pi(\theta|D) \) are available.

The advantages of the Bayesian paradigm over the frequentist paradigm in survival analysis are the following:

- Survival models are quite hard to fit, especially in the presence of complex censoring schemes, and MCMC techniques, fitting complex survival models is fairly straightforward.

- MCMC sampling enables to make exact inference for any sample size without resorting to asymptotic calculations, and in frequentist paradigm, for example, variance estimates usually require asymptotic arguments which can be quite complicated to derive and in some models simply not available.

- The Bayesian paradigm is a powerful tool for quantifying prior data, especially historical data.

More detailed discussion of the advantages of Bayesian paradigm and their demonstration can be found in [15]. Arbeev et al [8] present an example of implementation and comparison of traditional approach to the frailty model versus a Bayesian model using MCMC methods.
In the study we plan to apply Bayesian inference to the same problems considered in the previous sections and to compare the obtained results by this approach to the results obtained from a frequentist survival analyses.
Chapter 5

Retrial Analysis

The importance of the retrial analysis was described previously in Section 1. The aims of the research are as follows:

- to analyze customers’ retrial behavior in order to create more realistic models;
- to construct analytical models that can help us in performing the analysis of real systems;
- to identify problematic points in the management of a call center with retrials;
- to find ways for improving the performance characteristics (changing a service policy, adding ”monitoring stations”, etc.) by way of using analytical formulas and simulations.

This chapter presents some possible models that can be employed for such analyses.

5.1 Customer Retrial Behavior

To illustrate the customer retrial behavior, we analyzed the data for company C obtained during the period of a month. For the purpose of simplicity of the analysis, we limited the number of calls to those made by 170000 customers. These customers were randomly chosen from the data. Each of these customers called fewer than 30 times a month. This restriction was made in order to guarantee the genuineness of the customers (see 3.2).
5.1.1 Possible Call Scenarios

Each call has its own scenario (as described in Section 3.1). If the customer is served in the IVR mode and does not request the service of an agent, his/her call is classified as an "IVR only" call. We distinguish between an "IVR only" call and a call with a customer requesting an agent service. In our data, we have a special field named "call_outcome". This field has a number of codes, which we grouped and renamed to simplify the model. As a result, the following new codes for the "call_outcome" field were obtained:

1 - regular call, i.e. a call that was ended by the customer after receiving service;

2 - short abandonment call, i.e. a call that was ended by the customer within 5 sec. of the waiting period;

3 - abandonment call, i.e. a call that was ended by the customer after more than 5 sec. of the waiting period;

4 - system error call, i.e. a call that was ended due to some system error, such as a wrong transfer, an accidental termination, etc.;

5 - successful "IVR only" call, i.e. a call that was ended by the customer;

6 - unsuccessful "IVR only" call, i.e. a call that was ended because of a system’s error or was terminated by the customer for no special reason.

Figure 5.1 describes call outcomes for one-customer calls. The outcome of one call does not affect the next call outcome of the same customer. The transition probabilities are different for each state and can be calculated on the basis of the available data. The results presented in Figure 5.1 can be used in the simulation procedure.
5.1.2 Retrial Definition

Retrial analysis presented before (see [39], [38], [7]), deals with retrials after a blocking signal or after abandonment. We are also interested in the analysis of retrials after service. In [50], the authors deal with retrials that were triggered by unsuccessful service. In this research, the definition of retrials includes the calls of the same customers that took place within a short period of time after the primary call. At this point, the question may arise: "What is the time interval between a primary call and a retrial?" It may be expected, that a customer calling again immediately after receiving his/her service is not satisfied and would like to get a more comprehensive service. However, there are cases when the customer calls back after some extended period of time. For instance, a customer, who called in the morning, may decide to call back in the evening after work or even the next day. For us, this case is still considered a retrial. At the present state-of-the-art, there is no means to define the time lapse between the primary call and the retrial. Therefore, for the initial analysis, we define this interval as follows:

2 day-interval - the maximum period of time between the primary call and the next one, if a customer requested an agent’s service;

1 day-interval - the maximum period of time between the primary call and the next one, if a customer did not request the agent’s service and a primary call was "IVR only" call.
The above is an operational decision, and in the future we plan to provide a sensitivity analysis in order to check the results for the cases in which the definition of time between retrials is changed.

Figure 5.2: Customer retrial behavior scheme.

Figure 5.2 presents a scheme of customers’ retrial behavior. According to this scheme, a customer can make a number of attempts in order to receive a desired service. At attempt number $i \ (i = 1, 2, ...)$, a customer receives the service in the node $i$. Each attempt to receive service has one of the described possible outcomes. The call outcome has state $(i, j)$, where $i$ is a number of an attempt and $j$ is an outcome code. Then a customer can make another attempt (move to the next node), or he/she can end his/her current sequence of attempts and move to an idle state (state 0), and call again when needed. This scheme has some problematic issues. The first one is the transition from the call outcomes to another attempt. According to the scheme, a customer makes a new attempt in a new service pool. This is not true if a new attempt was started within a short time after the end of the previous attempt. In this case, the scheme does not describe a real situation, but we still use it because it is important for us to illustrate a customer’s behavior during the sequence of attempts. Another problem is the time scale. The time that passes between an idle state and the first attempt equals the time between the end of the sequence of attempts and the primary attempt of the next sequence. This transition can take weeks. The transition time between nodes and states describing the call outcomes, equals the queue exit time, i.e. the end of the service, the abandonment, etc. Such transition happens within minutes or even seconds. The transition between each of the call
outcomes and the idle state is immediate. Actually, this is an auxiliary transition, and it was created in order to illustrate the last call outcome. The transitions between the states of the call outcomes and the next attempt is a period of time till the next attempt. These time intervals can take seconds, minutes or hours. When creating analytic models, all these issues must be taken in account.

Our data allows calculation of customer behavior characteristics. For instance, the table below shows the probability of ending the sequence of attempts or making a new attempt, given the current attempt number and the call outcome.

<table>
<thead>
<tr>
<th>attempt</th>
<th>call outcome</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.616</td>
<td>0.193</td>
<td>0.001</td>
<td>0.008</td>
<td>0.003</td>
<td>0.116</td>
<td>0.062</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.379</td>
<td>0.214</td>
<td>0.050</td>
<td>0.068</td>
<td>0.008</td>
<td>0.117</td>
<td>0.164</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.331</td>
<td>0.361</td>
<td>0.007</td>
<td>0.089</td>
<td>0.007</td>
<td>0.089</td>
<td>0.116</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.285</td>
<td>0.372</td>
<td>0.014</td>
<td>0.047</td>
<td>0.019</td>
<td>0.168</td>
<td>0.095</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.612</td>
<td>0.042</td>
<td>0.000</td>
<td>0.002</td>
<td>0.000</td>
<td>0.290</td>
<td>0.053</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>0.471</td>
<td>0.114</td>
<td>0.001</td>
<td>0.007</td>
<td>0.001</td>
<td>0.137</td>
<td>0.268</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0.542</td>
<td>0.228</td>
<td>0.002</td>
<td>0.010</td>
<td>0.003</td>
<td>0.139</td>
<td>0.075</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.375</td>
<td>0.212</td>
<td>0.046</td>
<td>0.046</td>
<td>0.010</td>
<td>0.114</td>
<td>0.197</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.280</td>
<td>0.380</td>
<td>0.009</td>
<td>0.106</td>
<td>0.008</td>
<td>0.100</td>
<td>0.118</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.419</td>
<td>0.268</td>
<td>0.006</td>
<td>0.040</td>
<td>0.016</td>
<td>0.145</td>
<td>0.107</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.515</td>
<td>0.057</td>
<td>0.000</td>
<td>0.003</td>
<td>0.001</td>
<td>0.359</td>
<td>0.066</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>0.424</td>
<td>0.119</td>
<td>0.002</td>
<td>0.008</td>
<td>0.001</td>
<td>0.153</td>
<td>0.294</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0.485</td>
<td>0.253</td>
<td>0.002</td>
<td>0.013</td>
<td>0.003</td>
<td>0.156</td>
<td>0.088</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.294</td>
<td>0.259</td>
<td>0.040</td>
<td>0.070</td>
<td>0.010</td>
<td>0.124</td>
<td>0.204</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.220</td>
<td>0.405</td>
<td>0.013</td>
<td>0.121</td>
<td>0.016</td>
<td>0.095</td>
<td>0.130</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.332</td>
<td>0.311</td>
<td>0.021</td>
<td>0.026</td>
<td>0.013</td>
<td>0.153</td>
<td>0.145</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.463</td>
<td>0.061</td>
<td>0.000</td>
<td>0.003</td>
<td>0.001</td>
<td>0.397</td>
<td>0.076</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>0.376</td>
<td>0.120</td>
<td>0.002</td>
<td>0.008</td>
<td>0.001</td>
<td>0.161</td>
<td>0.331</td>
</tr>
</tbody>
</table>

Table 5.1: Transition probabilities of moving from the current outcome at a given attempt to the next state.

Table 5.1 presents the probabilities only for the first three attempts in order to illustrate the main trends.
5.2 Multiserver Queues

In this section we consider some theoretical models of a call center. Our goal is to calculate the performance measures of these models, and then, to find their approximations. Moreover, we are interested in the process simulation with the use of proposed models in order to check approximations and in order to analyze the changes in the behavior of the systems with changing parameters.

5.2.1 The Open Jackson Network

First, let us consider the following model for a call center: The arrival process of customers making a primary call, is a Poisson process with rate $\lambda$. There is an infinite number of trunk lines and $S$ agents in the system. Customers who find all servers busy join a queue. We suppose that there is a FCFS policy in the agents’ pool. If a customer is waiting in a queue, he/she may abandon the system after exponentially distributed time with rate $\theta$, or the call can be answered by an agent before this happens. Agent service time is described by independent identically distributed exponential random variables with rate $\mu$. Agent service time is independent of the arrival time. After receiving his/her service, the customer leaves the call center satisfied or unsatisfied with the service received. An unsatisfied customer may leave the system with probability $1 - p_R$, or return to the call center within a short period of time with probability $p_R$. After abandonment, a customer may redial with probability $p_{RA}$ within a short time, for instance 2 days, or leave the system. At this stage, we assume that a customer who has decided to recall after abandonment or on receiving unsatisfactory service, redials the call center after an exponentially distributed time with rate $\lambda_{RA}$.

The behavior of the system is described by a two-dimensional, continuous Markov chain $Q(t) = (Q_1(t), Q_2(t))$, where $Q_1(t)$ equals the number of customers residing in the service node (waiting or being served) and $Q_2(t)$ equals the number of customers in the retrial pool. Figure 5.3 presents the state transition diagram for the described system.
Figure 5.3: Markov chain for a queue with abandonments, retrials and revisits.

Actually, our model is a system with two multi-server queues connected in series. We can consider it as an open Jackson network, which is described in Figure 5.4.

Figure 5.4: The queue with abandonments, retrials and revisits.

For the described system, we can find a stationary probability and as a result, calculate performance measure expressions for the call center in question. In
the future, it would be interesting to find the approximations for performance measures for a system with a large number of servers and a large input rate. It would be also important to compare these approximations with the real call center data.

5.2.2 Time Varying Multiserver Queues

In practical terms, call centers are constantly subject to time-varying conditions. One of the possibilities to deal with this problem is to consider half-hour intervals and to assume that the parameters are constant and equal their means for each interval.

A more challenging approach would be to consider the system as time-varying multiserver queues. In this vein we introduce time-dependent rates for arrivals, abandonments and retrials, as well as a varying number of servers. Thus, we characterize our multiserver queue by the following parameters:

- \( \lambda_t \) - the external arrival rate for the calling node at time \( t \) \((\lambda_t \geq 0)\),
- \( \theta_t \) - the abandonment rate for the calling node at time \( t \) \((\beta_t \geq 0)\),
- \( \mu_t \) - the service rate for the calling node at time \( t \) \((\mu_t \geq 0)\),
- \( \lambda_{R_t} \) - the service rate for the retry node at time \( t \) \((\lambda_{R_t} \geq 0)\),
- \( p_{Rt} \) - the probability of retrial after a service at time \( t \) \((0 \leq p_{Rt} \leq 1)\),
- \( p_{RA_{At}} \) - the probability of retrial after abandonment at time \( t \) \((0 \leq p_{RA_{At}} \leq 1)\),
- \( S_t \) - the number of calling servers at time \( t \) \((S_t \in \{0, 1, 2, \ldots\})\).

The sample paths of the queue length process \( Q(t) = (Q_1(t), Q_2(t)) \) are uniquely determined by the relations:

\[
Q_1(t) = Q_1(0) + \Pi_{21} \left( \int_0^t Q_2(s) \lambda_{Rs} ds \right) - \Pi_{12} \left( \int_0^t (Q_1(s) - S_s)^+ \theta_s p_{RAs} ds \right)
+ \Pi_a \left( \int_0^t \lambda_s ds \right) - \Pi_b \left( \int_0^t (Q_1(s) - S_s)^+ \theta_s (1 - p_{RAs}) ds \right)
- \Pi_c \left( \int_0^t (Q_1(s) \wedge S_s) \mu_s ds \right)
\]

(5.1)
and

\[ Q_2(t) = Q_2(0) + \Pi_{12} \left( \int_0^t (Q_1(s) - S_s)^{+}\theta_s p_{RA_s} ds \right) - \Pi_{21} \left( \int_0^t Q_2(s) \lambda_{Rs} ds \right) \\
+ \Pi^d \left( \int_0^t (Q_1(s) \land S_s) \mu_s p_{Rs} ds \right), \]

(5.2)

where \( \Pi_{12}, \Pi_{21}, \Pi^a, \Pi^b, \Pi^c \) and \( \Pi^d \) are six given mutually independent, standard (mean rate 1) Poisson processes [35]. Here \( x \land y = \min(x, y) \) and \( x^+ = \max(x, 0) \) for all real \( x \) and \( y \).

As in [38], we consider the following asymptotic regime:

In a system with index \( \eta \), the only scaled parameters are the initial conditions \( Q_\eta^0(0) = [\eta Q_i^0(0) + \sqrt{\eta} Q_i^{(1)}(0)] + o(\sqrt{\eta}) \) for constants \( Q_i^0(0) \) and \( Q_i^{(1)}(0) \) (\( i = 1, 2 \)), the external arrival rate, i.e. the intensity of the Poisson arrival process is \( \eta \lambda_t \), and the number of servers is \( \eta S_t \). The scaled queue length process \( Q_\eta(t) = (Q_1^\eta(t), Q_2^\eta(t)) \) is then uniquely determined by the relations:

\[ Q_1^\eta(t) = Q_1^\eta(0) + \Pi_{21} \left( \int_0^t Q_2^\eta(s) \lambda_{Rs} ds \right) - \Pi_{12} \left( \int_0^t (Q_1^\eta(s) - \eta S_s)^{+}\theta_s p_{RA_s} ds \right) \\
+ \Pi^a \left( \int_0^t \eta \lambda_s ds \right) - \Pi^b \left( \int_0^t (Q_1^\eta(s) - \eta S_s)^{+}\theta_s (1 - p_{RA_s}) ds \right) \\
- \Pi^c \left( \int_0^t (Q_1^\eta(s) \land \eta S_s) \mu_s ds \right), \]

(5.3)

and

\[ Q_2^\eta(t) = Q_2^\eta(0) + \Pi_{12} \left( \int_0^t (Q_1^\eta(s) - \eta S_s)^{+}\theta_s p_{RA_s} ds \right) - \Pi_{21} \left( \int_0^t Q_2^\eta(s) \lambda_{Rs} ds \right) \\
+ \Pi^d \left( \int_0^t (Q_1^\eta(s) \land \eta S_s) \mu_s p_{Rs} ds \right). \]

(5.4)

Applying the Strong Law of Large Numbers and [35] we receive

\[ \lim_{\eta \to \infty} \frac{1}{\eta} Q_\eta = Q^{(0)} \text{ a.s.}, \]

(5.5)

\(^1\)Actually \( \eta S_t \) should be the integer part of \( \eta S_t \), but to avoid trivial complications and simplify notation, we assume it is just \( \eta S_t \).
where the asymptotic assumptions for the initial conditions are

$$\lim_{\eta \to \infty} \frac{1}{\eta} Q^\eta(0) = Q^{(0)}(0) \quad a.s.,$$  \hspace{1cm} (5.6)

and $Q^{(0)}(0)$ is a constant.

Moreover, based on [35], $Q^{(0)} = \{Q^{(0)}(t) \mid t \geq 0\}$ is uniquely determined by $Q^{(0)}(0)$ and autonomous differential equations

$$\frac{d}{dt} Q^{(0)}_1(t) = \lambda_t + Q^{(0)}_2(t) \lambda_{Rt} - \mu_t \left( Q^{(0)}_1(t) \land S_t \right) - \theta_t \left( Q^{(0)}_1(t) - S_t \right)^+, \hspace{1cm} (5.7)$$

and

$$\frac{d}{dt} Q^{(0)}_2(t) = \theta_t \rho R_A t \left( Q^{(0)}_1(t) - S_t \right)^+ - Q^{(0)}_2(t) \lambda_{RAt} + \left( Q^{(0)}_1(t) \land S_t \right) \mu_t \rho R. \hspace{1cm} (5.8)$$

To put it differently, $Q^\eta \approx \eta Q^{(0)}$ and we call $Q^{(0)}$ the \textit{fluid approximation} for $Q^\eta$. As shown in [36], [37] and [38], this approximation performs extremely well. Therefore, it is convenient to use it for the performance analysis of a call center with changing parameters.

It would also be interesting to refine this deterministic fluid approximation by deriving a stochastic \textit{diffusion approximation}. It is a sophisticated matter and we delegate it to future investigation.

\subsection*{5.2.3 Finite State Markov Chain}

Modern technologies allow increasing number of trunk lines according to the demand. Therefore, it is quite legitimate to suppose that this number is infinite. However, it does not mean that finite state models are not required. For instance, let us consider a specific kind of call center meant to serve some company’s customers. We can describe such call centers using so called subscriber models. In this case, we can limit the number of customers in the system by the total number of customers. We can model such call centers with the help of the process similar to the continuous Markov chain $Q(t) = (Q_1(t), Q_2(t))$ we referred to in Section 5.2.1. Let us recall that $Q_1(t)$ equals the number of customers residing in the service node (waiting or being served), and $Q_2(t)$ equals the number of customers in the retrial pool. The difference is that here we deal with a finite Markov chain. Figure 5.5 presents a transition diagram for this chain.
Figure 5.5: Markov chain for a queue with abandonments, retrials and revisits.

Such a representation gives us some tips for the analysis, namely, in finding the expressions for performance measures, it could be convenient to use a recursion algorithm.
5.2.4 The Closed Jackson Network

The system described in the previous section may also be presented as a closed Jackson network (see Figure 5.6).

Figure 5.6: Closed Jackson network for the queue with abandonments, retrials and revisits.

Such a representation helps us express the stationary probabilities and, as a result, performance characteristics. However, due to the complex nature of these expressions and the numerical instability associated with the computation process, the whole procedure may be time-consuming and ultimately produce inaccurate values. On the other hand, it is possible to use approximations for the system characteristics as in [31]. These approximations are convenient for the investigation of the effect of changes in the system parameters on the system performance. For instance, the increase in the probability of post-service redial can worsen the service in a call center, i.e. it can lead to increased waiting time and number of customers in the queue, and, as a result, an increased rate of abandonment.

Also, it is important to decide whether we are going to consider an open network or a closed one. On the one hand, an open Jackson network makes
the analysis easier. On the other hand, in some cases a closed network is more suitable, because, for instance, it excludes the possibility of explosions.

5.2.5 Comprehensive Model

In addition to providing agent service, a modern call center commonly supports an interactive voice response (IVR) service. Thus, the model considered in Section 5.2.1 can be extended to include a new node reflecting IVR service. This IVR node involves a $M/M/\infty$ queue model.

According to this extended model, at the beginning of the call, a customer receives service from the IVR processor. The assumption is that the IVR processing time intervals are independent and they represent identically distributed exponential random variables with rate $\delta$. Then, the customer faces the choice of leaving the system ($1 - p_S$ probability) or requesting agent service ($p_S$ probability). In our previous paper [31], we proposed a model similar to that of Section 5.2.1, though it did not support possible customer retrials.

The current model is constructed taking into the account the information regarding call outcomes that were presented earlier in Section 5.1. The retrial node from the original model was replaced by three nodes per each outcome, namely, abandonment ($Ab$), post-service ($R$), and call error ($E$) retrials.

![Figure 5.7: Open Jackson network for the queue with abandonments, retrials and revisits.](image-url)

49
Figure 5.7 presents an open Jackson network that reflects the current model. The proposed analysis of previous models is relevant to the current one, but here it would be quite complicated. It is also possible to consider our model as a closed Jackson network. However, this will make the model even more complicated. On the other hand, both models could allow us to obtain performance characteristics that could contribute to formulating solutions, thereby essentially improving the service level of the system.

5.3 Additional Retrial Analysis

In the two previous sections, we described possible analytical models that could be helpful in the calculation of performance measures, in creating simulations and the analysis of systems with changing parameters. Obviously, real call center data can improve these models by allowing the estimation of model parameters and testing approximations. However, the opposite process is also possible. An example would be the use of an analytical model to compensate for the lack of information. Our original data show the total arrival rate, and we are unable to distinguish between retrials and primary calls because these data may contain some fake customer identity numbers (see Section 3.1). What is the way to estimate retrials for calls without a customer I.D.? Here, theoretical models could be of help.

For a call center manager, it is very important to be able to assess professional qualities of agents. The following parameters may be considered: the accuracy of the information provided, its comprehensiveness, the ability of the agent to identify real needs of his/her customer and the tact with which the agent fulfills his/her duties. There are some widely used ways of providing such quality assessment, which include call recording, selective listening and customer polling. These procedures require additional staffing and may be quite complicated for big call centers. Therefore, we would like to develop an algorithm for identifying possible problematic issues. For this purpose, we can use the phenomenon of post-service retrials that usually signal customer dissatisfaction with the service. On the basis of after-call information, we can build a grading system that will process the number of retrials and the total time spent on these repeated calls. In practical terms, the proposed algorithm can help cut the time consumption and expenses for the quality assessment system and ultimately increase its effectivity.
Chapter 6

Research Goals

This study is motivated by call center management problems and the opportunity to analyze a large quantity of data collected over a long time period (more than 3 years). The main idea is to apply methods of statistical analysis to call center data in order to identify basic problems induced by abandonment and retrial phenomena, to find the sources of such problems, to develop ways for their solution and to estimate their possible impact. This chapter summarizes and specifies the goals of the ongoing research as follows:

1) Characterize customer patience by using real call center data and methods of survival analysis and fit the models of survival analysis that were proposed in Chapter 4. To compare the fitted results and relevance of models used in this analysis.

2) Construct procedures for checking the dependence structure of clustered failure-time data in the hierarchical models that are proposed in Section 4.1.2.

3) Fit the survival analysis models with stochastic process frailty for our data: to estimate the diffusion parameters and to construct an appropriate hazard function for models with frailty as a Lévy process.

4) Develop an appropriate first-passage-time model to our data, apply it and compare the results with outputs of the survival analysis models.

5) Consider the use of copula models and for the analysis of our data.

6) Apply Bayesian inference to the problem of the shared frailty model and the hierarchical frailty model, and compare the results.

7) Analyze customer retrial behavior in order to create the most reliable models of this phenomenon, find an optimal definition for the retrials using sensitivity
analysis and determine the agent service grade (discussed in Section 5.3) by analyzing call center data and examining possible scenarios with the help of simulation of theoretical models.

8) Provide a theoretical investigation of performance characteristics for appropriate models and to compare the theoretical results with output of simulations and real data analysis.

9) Develop an algorithm for identification of the problematical issues in an agent’s service and to construct an optimal policy for minimization of percent of abandonments and retrials and test the results by simulation.
Bibliography


54


