Stochastic Modeling and Management of an Emergency Call Center

A Case Study at the Swedish Emergency Call Center Provider, SOS Alarm Sverige AB

Klas Gustavsson

Department of Information Systems and Technology
Mid Sweden University

Licentiate Thesis No. 141
Sundsvall, Sweden, 13 juni 2018

Mittuniversitetet
Informationsteknologi och medier
SE-851 70 Sundsvall
SWEDEN

ISBN 978-91-88527-58-5
ISNN 1652-8948

Akademisk avhandling som med tillstånd av Mittuniversitetet i Sundsvall framlägges till offentlig granskning för avläggande av teknologie licentiatexamen

Onsdagen den 13 juni 2018 i L111, Mittuniversitetet, Holmgatan 10, Sundsvall.

Tryck: Tryckeriet Mittuniversitetet

To My Wife
To Ebbe

Abstract

A key task of managing an inbound call center is in estimating its performance and consequently plan its capacity, which can be considered a complex task since several system variables are stochastic. These issues are highly crucial for certain time-sensitive services, such as emergency call services. Waiting times affect the service quality of call centers in general, but various customers may place different waiting time expectancies depending on the need. Call center managers struggle to find the relationship between these expectations to their strategical, tactical and operational issues. They are assisted by queueing models that approximate the outcome. Simple setups use analytical approximations while a network of multi-skilled agents serving several customer classes is dependent on computer simulations. Regardless of simple or complex setups, models assume that the system components are homogenous, that the components have some parametric distribution, and that they remain the same regardless of the setup. Human resource and marketing research show that such status quo assumptions are not highly reliable. As an example, customer experience is often affected by the skill of the agent, and agents themselves are affected by their workload and duties, which inter alia affect their efficiency. This thesis aim to assist the Swedish emergency call center with a strategical issue, which require detection of some causalities in the set of system components. The overall aim is to design a simulation model, but such model requires a lot of detailed system knowledge, which itself adds to the knowledge gap in the research field. Findings that contribute to the scientific knowledge body include the burst model that addresses some of the non-stationarity of call arrivals, since some rapid rate increments derives from a latent emergency event. Other contributions are the introduction of stochastic agent behavior, which increases the uncertainty in queueing models; and the service time relationship to geographical distance. The latter may involve general evidence on how area-specific understanding and cultural differences affect the quality of service. This is important for organizations that consider off-shoring or outsourcing their call center service. These findings, along with several undiscovered and unknown influencers, are needed in order to design a reliable simulation model. However, the proposed model in this study cannot be rejected, in terms of waiting time replication. This robust model allowed traffic routing strategies to be evaluated and also assisted managers of the emergency call center into a strategical shift in the late 2015.

Acknowledgements

I would like to take the opportunity to thank a number of important individuals who have supported me throughout this journey. Firstly, my supervisor Leif Olsson has played a key role in my studies, as he guided me through a dense jungle of theories and research orientations at the initial stage. Also, my associate supervisor Mikael Gidlund and my mentor Erik Borglund need to be mentioned.

My research would not have been possible without the assistance from the people at SOS Alarm. The most important driver for this project is Ulf Andersson. Both Ulf and Erik Borglund came up with the idea of this research collaboration over a night of wine. Ulf has always fought to give me opportunities to conduct research. My fondest memory of his is probably his sales skills during a poster session at INFORMS Nashville 2016. It is highly likely that there has never been a time where so many were coerced to look at a poster. Another person who has been instrumental to his project is Henrik Alm, who for various reasons has assumed the baton stick during the last year. I would also like to extend my gratitude to the team at SOS Alarm, namely Claes Eliasson, Christine Stadling, Eva Bekkevik and Mikael Björkander, who are absolutely critical actors in this project.

I am very grateful to Pierre L'Ecuyer, who welcomed me to Montreal and deepened my knowledge about stochastic modeling. From our campus, there are also a lot of individuals who influenced me. A special thank you must be addressed to Kristoffer Karlsson at the Department of Mathematics and Science Education, who has assisted me with solutions and understandings in a complex mathematical domain. I would also like to thank my informal supervisor Aron Larsson, especially for his help during conferences; he is also an intelligent creature with whom you can have fruitful discussions. I also must take this opportunity to thank my floor mates Christine Grosse and Leif Sundberg for always being willing to discuss practical and theoretical issues.

As seen above, there are many important people that I feel a need to show my absolute greatest appreciation to. However, the most essential ones are my family. My father, mother and sister have always been a source of encouragement for me. Lastly, the most important person who has been with me during this journey is my wife Marlene who always encourages me; and most importantly, she is my stress manager. Marlene is a very intelligent person in many ways, and I have never ceased being surprised by her.

Contents

Abstract **v**

Acknowledgements **vii**

List of Papers **xiii**

Terminology **xix**

1 Introduction **1**

1.1 Call center as a service 2

1.2 Call center as a queueing system 2

1.2.1 Poisson process 4

1.3 Call center operations and management 4

1.4 Research gaps in call center operations and management 5

1.4.1 Call center design 5

1.4.2 Arrival process 6

1.4.3 Multi-disciplinary research 7

1.5 The case: An emergency contact center 7

1.5.1 Call center design 8

1.5.2 Call assignment 10

1.5.3 Research rationale 10

1.6 Overall aim and objectives 10

1.7 Decomposed and verifiable goals 11

1.8 Academic contribution 12

2 Theory **15**

2.1 Call center operations management . . . 15
2.1.1 Operational regime . . . 15
2.1.2 Workforce management . . . 16
2.2 Service time . . . 17
2.3 The arrival process . . . 18
2.3.1 Time dependence . . . 18
2.3.2 Overdispersion . . . 19
2.3.3 Intra- and interday dependencies . . . 19
2.3.4 Latent variables dependencies . . . 20
2.3.5 Forecasting . . . 20
2.4 Abandonments (customer patience) . . . 21
2.5 Agent efficiency . . . 21
2.6 Capacity estimation . . . 21
2.7 Multi-skill and pooling call centers . . . 23
2.8 Simulation . . . 24
2.8.1 Random number generation . . . 25
2.8.2 Discrete-event simulation . . . 26
2.8.3 Agent-based simulation . . . 26

3 Methodology **27**
3.1 Design science research . . . 27
3.2 Philosophical standpoint . . . 28
3.2.1 Research perspective . . . 28
3.3 Research process . . . 29
3.3.1 Paper I . . . 30
3.3.2 Paper II . . . 30
3.3.3 Paper III . . . 31
3.4 Data collection . . . 32
3.5 Ethical consideration . . . 32

4 Model **35**
4.1 Variable assessment . . . 35
4.1.1 Arrival process . . . 35
4.1.2 Agent . . . 36

4.1.3 Customer 36
4.2 Schematic description of the model 37

5 Results **39**
5.1 Call arrival process 39
5.1.1 Burst phenomenon 40
5.2 Service time 44
5.3 Agent behavior 45
5.4 Simulation model validity 46
5.5 Routing and pooling effects 47

6 Discussion **49**
6.1 Routing effects observed empirically 49
6.2 Research limitations 50
6.3 Future work 51
6.3.1 Cross-disciplinary approach 51
6.3.2 Implications of exploiting bursts 52

7 Conclusions **53**
7.1 Research findings 53

Bibliography **55**

Biography **59**

List of Papers

This thesis is mainly based on the following papers, herein referred by their Roman numerals:

I K. Gustavsson, P. L'ecuyer and L. Olsson. Modeling bursts in the arrival process to an emergency call center. Proceedings of the 2018 Winter Simulation Conference. [Submitted]

II K. Gustavsson. Service time effects of distancing from the customer, a case study from the Swedish emergency call center. 2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM).

III K. Gustavsson, L. Olsson, M. Gidlund and U. Andersson. Simulating routing strategies of inbound calls at the Swedish emergency call center. European Journal of Industrial Engineering, 2018. [Submitted]

List of Figures

1.1 Call centers represented as a queueing system (Gans et al., 2003). . . . 3

1.2 Inhomogeneous Poisson process with piecewise constant $[t_1, t_5]$, linearly increasing $[t_5, t_6]$ and exponentially decreasing intensity $[t_6, t_7]$. . 5

1.3 Overview of the emergency center mission and their key stakeholders. The thicker arrows represent the communication to the public, while the solid lines represent communications channels to societal resources, and the dotted lines indicate other stakeholders who put expectancies and requirements on the service. 8

1.4 Map of counties in Sweden including the location of emergency centers as well as strategic links between counties and primary offices and regions. The outlined geographical mapping is hereafter referred to as agent and customer classes. 9

1.5 Considered traffic routing strategies evaluated in this study. The grey nodes correspond to the initial SBR, and the following nodes are the next pooling stage. The 13 top-circles are the different sites located in Sweden divided into three regions of collaboration, and the bottom circle is a national collaboration queue. The geographical mapping plus the sites that belong to each node is found in Figure 1.4. 11

2.1 Literature review of existing rate models. 18

2.2 Canonical representation of some common designs of skills-based routing (Garnett and Mandelbaum, 2000). 24

2.3 The different types and combinations of networking designs (Gans et al., 2003). 24

2.4 Example of Archimedean copulas with $n = 10000$ and $\theta = 7$. 26

3.1 DSR cycle according to Owen (1998). 27

4.1 Overview of SBR implementation in the simulation model. 37

4.2 Overview of the dynamic overflow/pooling setting implemented in the simulation model, together with stochastic and deterministic assumptions. 38

5.1 Box-plot of arrivals during 15 minute intervals during hours and days. Data from January-June 2016. 39
5.2 Variability within assumed stationary 30 minute intervals, calculated as subset m rate divided by its corresponding interval j rate. 40
5.3 Cumulative arrival count of empirical arrivals to the call center compared to expectancy of assuming a stationary period. 41
5.4 Cumulative count of arrivals that reports the same event. 41
5.5 Empirical dependencies between parameters A, B &C 42
5.6 Empirical dependencies of the CDF of parameter A, B and C 43
5.7 Empirical service time PDF (2016). 44
5.8 PDF comparison between empirical waiting times and corresponding simulation output. 45
5.9 CDF comparison of the evaluated routing strategies. 46
5.10 Empirical service time of county (customer class) and site (agent class) 47

List of Tables

3.1 Research overview 30

5.1 Operator absence modeled in Arena as failures with up time corresponding to time between absences and down time to the time being absent. 45

5.2 Full table of routing strategy performance based upon 15 simulation replications, n=33318. 47

6.1 Empirical call allocation during September 2016 to August 2017 50

Terminology

Abbreviations and Acronyms

ABS	Agent-Based Simulation
ACD	Automatic Call Distributor
AHD	Average Handling Time
ASA	Average Speed of Answer
CDF	Cumulative Density Function
CRM	Customer Relationship Management
CTI	Computer-Telephone Integration
DES	Discrete-Event Simulation
DSR	Design Science Research
ED	Efficiency-Driven (operational) regime
FIFO	First-In-First-Out (queueing discipline)
HR	Human Resource
IVR	Interactive Voice Response
OR	Operations Research
PDF	Probability Density Function
QED	Quality-Efficiency-Driven (operational) regime
QD	Quality-Driven (operational) regime
SBR	Skills-Based Routing
SLA	Service Level Agreement
TSF	Telephone Service Factor

Mathematical Notation

ρ	Agent occupancy
λ	Call arrival rate
μ	Service rate per server ($\mu = E[S]^{-1}$)
$C_\theta(u, v)$	Copula describing the joint distribution, i.e. dependence structure, between vectors u and v, with parameter θ
θ	Copula parameter, deciding the strength of dependence structure

Chapter 1

Introduction

This thesis deals with issues that concern call center managers in general, and particularly this work would be of interest to managers of emergency call centers. Issues concerning production planning and strategical routing are addressed, as well as important stochastic variables that affect the system. The research and its direction is a combination of building application-driven proposals and filling academic knowledge gaps. Selected directions mainly derived from case-specific characteristics of the Swedish emergency call center provider, SOS Alarm Sverige AB, who has funded this research project and is also providing important information regarding the service.

Call center services lies in a broader discipline referred to as service science, management and engineering. Maglio and Spohrer (2008) defined a services system as the "value-co-creation configurations of people, technology, value proposition connecting internal and external service systems, and shared information (e.g., language, laws, measures, and methods)." With this in mind, service science is the study of service systems that combines human, business, organization, and technological perspectives. Service systems have numerous issues that span over multiple research disciplines. This study uses the perspectives and approaches from the fields of operations research (OR), telecommunication, and queueing theory. The greatest common thread throughout this thesis is to understand the service expressed as a queueing system, which focuses on the technological perspective and the assumptions contained therein. Both the scope and consequently the system boundaries are determined by the call center representation as a queueing system (see section 1.2). Therefore, there is much emphasis placed on understanding the variables affecting the queueing system. Since many of these variables are random, much of the focus is on stochastic modeling.

1.1 Call center as a service

Mehrotra (1997) focuses on a technological perspective of service science and defined call centers as "any group whose principal business is talking on the telephone to customers or prospects." Batt et al. (2009) uses a Human Resources (HR) perspective and define call centers as "organizations that manage customer service and sales transactions across a wide range of product markets". Feinberg et al. (2002) uses a marketing perspective and stated that "call centers allow a company to build, maintain, and manage customer relationships by conducting transactions, giving information, answering questions, solving problems and resolving complaints quickly, and less expensively than face to face contact".

The service and customer value is created in the communication between customers and the individuals who serve customers. These individuals are known as agents in this thesis; specifically, agents refer to people scheduled by the call center to talk to customers. The contact to a call center can both be customer-initiated (inbound), agent-initiated (outbound), or a mixture of both. From a manager perspective, there are several non-trivial planning issues, such as dimensioning the call center. In an outbound service, such a decision is more trivial due to the fact that there is a simple causal relationship between capacity and productivity (i.e., agents and their workload). The productivity in such services is often based on measures derived from either the calls made or successive calls. In such services, the work load is system deterministic, meaning that either the system or the agents decide the workload. On the other hand, an inbound call center has a stochastic workload, as it depends on variables not within system control. In such a case, there are no simple relationship between workload and capacity. The workload is no longer deterministic but stochastic, since it depends on the call rate variation. When all agents in a call center are busy, customers are either blocked or placed in a queue. The waiting time is also an important part of the value-creation that a call center provides to its customers and is often the metric emphasized in capacity planning (Koole and Pot, 2006). Depending on the service that a call center offers, customers may have different expectancies regarding the service and their willingness to wait.

1.2 Call center as a queueing system

Call centers consist of k trunk lines that connect a call to the center. These trunk lines are occupied by $w \leq k$ workstations at where there are $m \leq k$ agents scheduled. When a call arrives, one of three scenarios could take place: (a) the call is distributed directly if there are idle agents, (b) the call is placed in a queue if there are no idle agents, or (c) the call is blocked if there are no trunks available. This is known as the general queuing representation (see Figure 1.1). In such a system, the resources are immediately released once a call has been served. There are a few types of arrivals (or callers) to such a system, namely those who make an initial contact, those who redial due to a busy signal, and those who abandon the queue upon waiting. A fourth example is customer who returns after talking to an agent, perhaps due to technical

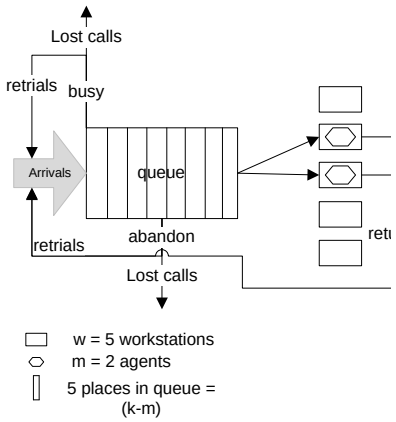


Figure 1.1: Call centers represented as a queueing system (Gans et al., 2003).

failure. It is important to distinguish between these types of arrivals, as they have a causal relationship to the state of the system. An explanatory example would be that the system cannot expect any retrials due to blocking or queueing when the system is relaxed (i.e., when agents are available). On the contrary, a higher number of retrials are expected during busy hours when the system is strained (i.e., during long waiting times). Kendall (1953) invented the Kendal notation that described the queue by three factors. The system has afterwards been extended with an additional three factors. The Kendal notation is today generally expressed with the following six factors and queue properties, represented as $1/2/3/4/5/6$, where each numbers position represent:

1. Arrivals process (generally Markovian)
2. Service time distribution (generally Markovian)
3. Number of servers
4. Number of trunks (infinite if left out)
5. Size of calling population (infinite if left out)
6. Queues discipline, generally First-In-First-Out (FIFO)

The simplest queue is the $M/M/1$ queue, which is composed of the Markovian arrival process, a random exponential distributed service time, and one server. However, such a model is most often an over-simplification, as seen with the number of possible retrials in Figure 1.1 (Gans et al., 2003).

1.2.1 Poisson process

Arrivals are the main stochastic driver in many service systems. The process of call arrivals are generally modeled as a Poisson process, which is a counting process $\{N(t), t \epsilon [0, \inf)\}$ with rate λ and the following properties (Saaty, 1961);

1. $N(0) = 0$,
2. $N(t)$ has independent increments,
3. $N(t) - N(s) \sim Poisson(\lambda(t-1))$, for $s < t$.

These properties mean that during a homogenous period (i.e., constant intensity λ), the inter-arrival times in a Poisson process are independent and exponential distributed. The random variable expressing the number of arrivals up to time t is $N(t)$ with an expectancy of $Poisson(\lambda t)$. Properties of the Poisson process is that $N(t)$ has random independent increments. The Poisson process inherits important statistical properties that simplifies the estimation of required agents to control quality metrics such as the average speed of answer (ASA) and the telephone service factor (TSF), which measures the fraction of calls answered within a fixed time threshold.

However, the intensity is not always constant, and the arrival process is not always homogenous. It can be considered a radical simplification to assume a constant rate for the center regardless of the time of day, week, and year. In such cases, the intensity is a function of time t, $\lambda(t)$. The intensity can sometimes be expressed as a piecewise constant function, and these are modeled as a nonhomogeneous Poisson process; examples can be seen in the periods $[t_1, t_4]$ and $[t_4, t_5]$ in Figure 1.2.1. This is relatively easy to implement as separated homogeneous processes, and this is often applied in call centers due to scheduling and forecasting advantages. There are also periods with non-constant intensity (see period $[t_5, t_7]$); such an intensity is generally not applied in call centers. The reason for this is difficulties in estimating such an intensity; because only arrivals are observed and not the intensity. In other words, it is difficult to distinguish statistically expected variability from inhomogeneity. Another reason is the difficulty in connecting such varying load to a reasonable staffing policy.

1.3 Call center operations and management

Many operational issues of call centers can be traced back to the design of the queuing system. The overall aim of these models is to allow managers to operationalize their efficiency and service quality trade-off. A call center system is generally built on several components, namely a computer–telephone integration (CTI), an automatic call distributor (ACD), a customer relationship management (CRM), and sometimes an interactive voice response (IVR). The CTI allows the system to extract information from the caller to the CRM; an example of this information would be the origin and number of the call. Such an extraction of information is a vital part in virtual

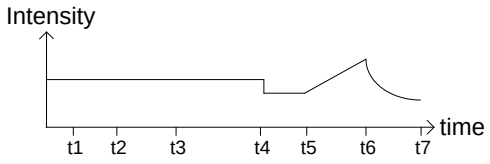


Figure 1.2: Inhomogeneous Poisson process with piecewise constant $[t_1, t_5]$, linearly increasing $[t_5, t_6]$ and exponentially decreasing intensity $[t_6, t_7]$.

call centers that use skills-based routing (SBR). SBR uses the ACD to route calls to an appropriate agent. It should be possible to extract some key metrics from the system; these metrics are important management tools in strategical and operational issues. Many OR models within call centers depend on statistic properties of queueing theory. Generally, call center models use system primitives of arrival rates, service time, and abandonment behavior to estimate the performance in terms of waiting times and abandonment rates (Gans et al., 2003). To obtain reliable results, models need to have detailed and reliable data, which is often difficult to obtain (L'Ecuyer, 2006). System primitives are usually simplified to meet assumptions in theoretic models. Gans et al. (2003) highlighted this issue and stated that at the time of their research, there has yet to be stochastic variables and causality that have been found in the call center context. Since then, there has been much advancement, and while some issues are resolved, there are still numerous knowledge gaps to fill. A new call center survey in 2007 pointed out that strategic decisions of multi-site operations have not been addressed properly in OR literature. Instead, the focus of call center OR during the last decade has shifted to exploit new technological features—such as IVR, designing new scheduling and planning models, and forecasting methodologies.

1.4 Research gaps in call center operations and management

1.4.1 Call center design

Gans et al. (2003) conducted an extensive survey of existing knowledge, technology, and research prospects regarding call centers. They addressed a lot of issues that had not been sufficiently covered or resolved in up-to-date literature. For instance, they stated that there has been little investigation on call center networking, including both quantitative effects based on queueing theory logics and the impact on actors involved. Research focusing on multi-site or multi-skill operations has often shared the original foundations of telecommunication and queueing theory. Gans et al. (2003) stressed the complexity of dynamic load-balancing and overflow protocols; they also emphasized how these designs jeopardize Poisson assumptions, such

as stationarity and one arrival at a time. The way these designs impact system planning has not yet been addressed, and it is necessary to find a systematic approach that captures the behavior of such schemes in order to understand the advantages and disadvantages.

1.4.2 Arrival process

A success factor for adequate decision making is modeling the stochastic variables and processes correctly. Understanding the arrival process is crucial in performance estimation and capacity planning. Even though models generally assume that arrivals is a Poisson process, studies have shown that key assumptions are often violated (L'Ecuyer, 2006; Jongbloed and Koole, 2001; Soyer and Tarimcilar, 2008). An important property of Poisson process is that callers decide to call independently of one another (Cinlar, 2013). Such an assumption may be questioned. The number of arrivals within a time segment depend on numerous variables. Some are periodic and predictable; some are unpredictable, and some are a stochastic process themselves. Periodic variables are relatively easy to understand and are accounted for in forecasting, whereas the unpredictable variables are more difficult to assess and understand because they are random in both time and magnitude. There are two issues that make this assessment non-trivial, namely one that identifies appropriate time segments in which the rate is constant and one that forecasts the expected number of arrivals within a particular time segment. The latter is a relatively straightforward task with practices that are applicable in a wide spectrum of fields; the former is more complex since the interval length is a random variable, which is not observable in call center data. Ontologically, the interval length at which the rate is constant depends on variables that can affect an individual's willingness to call; only some variables may be predictable. The duration and magnitude of these variables is complex to determine due to the fact that they are random variables. In other words, arrivals depend on some underlying space consisting of predictive and unpredictable variables that affect an individual's willingness to call. Consequently, each arrival is related to some underlying set of variables. Within telecommunication, despite predictable variables, the relations between points of arrivals are frequently neglected. Such neglect leads to an underestimation of the variability of the rate. This is especially the case in an emergency center context, where several people can be affected by a certain event. The accompanying arrival peak is normally not accounted for within capacity planning; the peaks are generally smoothed out in aggregations. Hereafter, we will call these rate jumps *bursts*. With the ontological view, the underlying event that triggers a call is a variable that affects one or several individuals' attitude towards making a call. The latent event adds a kind of relation between arrivals since the arrival occurred due to the same reason.

The understanding of bursts is limited within call center research, but the phenomenon needs to be addressed in order to understand how it performs and why the variability is greater than what the Poisson process suggests.

1.4.3 Multi-disciplinary research

The most recent survey on call center research was conducted by Aksin et al. (2007), and they highlighted the need to include sociological parameters (i.e., behavioral issues) that affect call center operations. Similarly, Gans et al. (2003) called for multi-disciplinary research to address some of the interdependencies between system components. For instance, several factors can highly affect the performance of a service, and these factors include agent preferences, their attitude, and how those variables vary during a shift. Such dimensions are advantageously translated into service time and efficiency, but there is a lack of knowledge on the underlying causalities. Customers' preferences and dynamics are also of importance, and these becomes more obvious when considering other quality of service measures. Aksin et al. (2007) stressed that service quality metrics–aside from those derived from waiting times–need to be included in the queuing system in order to obtain a representative view of the problem. Qualitative service metrics such as agent competence, politeness, and friendliness have also been shown to affect customer satisfaction, although it is not yet known what exactly the causation is between these variables and management decision in call centers. To complicate it even further, the strategical and operational decisions will likely affect the long-term attitude and well-being of the agents. In the long run, managerial decisions may be seen to have some impact on agent incentive, turnover, sickness, and absenteeism levels. These parameters have been stressed by Batt (Batt, 2002; Batt and Moynihan, 2002), but their dependencies to operational decisions has not yet been included in OR research using queuing theoretic models. For this reason, the telecommunication perspective of making rational management decision could be questioned (Aksin et al., 2007).

1.5 The case: An emergency contact center

Emergency contact centers are a type of call center. The service has a strict expectancy of waiting times and quality of service because the difference between life and death may be determined by a matter of seconds. Emergency centers are vital to any society as they provide the first assistance when individuals or property are threatened. Emergency call agents use telecommunication to guide and coordinate information and actions to assist citizens in need and the required emergency units. In Sweden, the organization SOS Alarm provides the emergency 112-service, and the regulation of the service is described in an agreement with the Swedish government. Although the Swedish emergency service is managed as a privately held unit, there are only two shareholders: the government and the Swedish Association of Local Authorities and Regions (SALAR). The intensity of Swedish inbound 112 emergency calls depends on various stochastic variables and has periodic patterns depending on season and day of the week. Between 2015 and 2016, there were roughly 60,000 calls a week, or 7,000 to 12,000 calls a day. Approximately 10% of the calls are unanswered for various reasons.

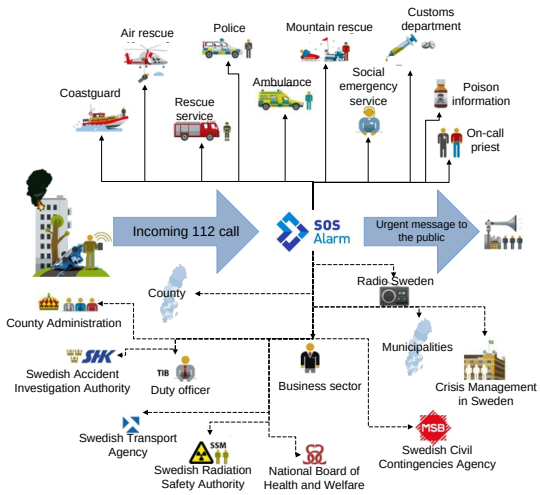


Figure 1.3: Overview of the emergency center mission and their key stakeholders. The thicker arrows represent the communication to the public, while the solid lines represent communications channels to societal resources, and the dotted lines indicate other stakeholders who put expectancies and requirements on the service.

1.5.1 Call center design

In Sweden, the emergency service is divided into 22 different counties, and there are 13 physical emergency centers (see Figure 1.4). SOS Alarm has classified agent and customer classes depending on that geographical mapping so that each county has a primary center where generally all dispatches are performed. Since the queueing system is a virtual call center, it is possible for customers in any county to be routed to any site. So the *county-to-site mapping* is one of the difficult planning decisions for managers to take.

Generally, each site has agents that perform three main tasks:

- Provide first assistance and assess the needs of all inbound 112-calls
- Strategic coordination and allocation of care resources, such as ambulance vehicles
- Coordinate rescue units and act as a link between the parties involved.

The three main duties translate into functions vacated in each emergency center, and agents can have a number of different roles during their scheduled hours. Most agents service inbound 112-calls, while a smaller number of agents are ambulance or rescue dispatchers. Coordinating ambulance and rescue operations are frequently less intense; hence, some of the dispatchers also support the 112-service when time

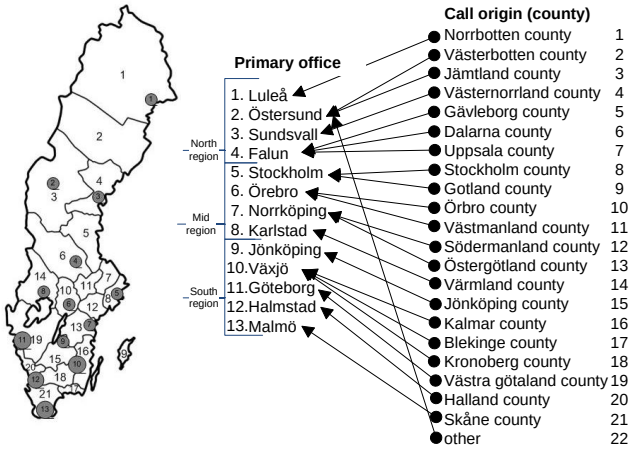


Figure 1.4: Map of counties in Sweden including the location of emergency centers as well as strategic links between counties and primary offices and regions. The outlined geographical mapping is hereafter referred to as agent and customer classes.

permits. Agents who primarily focus on other duties are referred to as *multi-duty agents*. This study only focuses on 112-service.

When this study was initiated, inbound emergency calls were routed to one of the centers in the region from where the call originated, as geographic proximity is considered beneficial (see north, mid, and south regions in Figure 1.4). Consequently, each region needed to balance their capacity to cope with their call load. The definition of a region and the connection between county and office is outlined in Figure 1.4. The location in combination with the different duties at each office form numerous routing capabilities. Routing determines the appropriate agent to a customer.

An issue that SOS Alarm currently faces is with deciding the routing scheme of arriving emergency calls depending on the county of the customer. Estimating alternatives performance is a non-trivial task due to the growing number of states that the system can assume as routing is utilized. The considered strategies uses three levels of centralization, through which the aim is to minimize the utilization of the centralization but with the contradictory objective of being effective and meeting Service Level Agreements (SLAs) based on TSF and ASA. The three levels of centralization is expressed as *agent classes* with different skills:

1. Primary agent class is the best matching agent class based on the geographical mapping presented in Figure 1.4.

2. 2. Secondary agent class is a match between agent site and a customer calling from an extended region, specifically the combined counties and agent classes (see regional division in Figure 1.4).

3. Tertiary agent class is any other agent than primary and secondary agents.

1.5.2 Call assignment

When learning about the system, one challenging task is to understand the call assignment feature of the ACD. SOS Alarm uses an agent-controlled assignment, which leaves the responsibility of picking a call to agents themselves. The feature benefits from a flexibility that allow agents to make situational decisions, such as during heavy traffic where there could be multiple calls for the same event or when there are calls identified as malicious, which should be prioritized accordingly in the system. However, the feature hampers the FIFO assumption, and it disables conventional routing methods where the system detects which agent is most appropriate based on some criteria and subsequently assigns the call to that agent. This would be possible in a push functionality when an agent has a status that is either idle, occupied, or blocked. Idle means that agent is available, and the others mean the agent is not available. In a pull system, the agent is either occupied, blocked, idle, or available, or the agent is simply idle and unavailable. System characteristics and its pull system add the stochastic parameter unavailability of the agent; since it is not possible to track agent statuses in the current system, the assessment of these variables is a challenging task.

The system implements *SBR* as different agent views of picking pools, which allows different agent classes to answer a call. The absence of the push functionality forces the system to use a threshold value before allowing other agents to answer. The overflow setting implies a dynamic routing scheme with cross-trained agents.

1.5.3 Research rationale

In 2015, SOS Alarm Sverige AB was considering alternative traffic routing (see Figure 1.5). Both the complex routing and random variables have made it difficult to predict the outcome and differences between the strategies. Up to now, there is limited information on the complete effects of complex routing modifications, especially regarding how to evaluate a dynamic overflow and pooling schemes, which is an issue that has not yet been presented in current call center research.

1.6 Overall aim and objectives

To analyze and assist the call center of SOS Alarm, the study uses the standpoint of OR and the assumptions within telecommunication and queueing theory. The overall aim is to optimize a multi-agent and multi-customer system, with the objective to maximize the complex utility of having an agent close to the customer location while not compromising neither SLAs nor resource costs. The system boundaries are decided by the queueing theoretic perspective, but they are also extended with stochastic resources. To achieve the overall goal, it is necessary to adapt the design

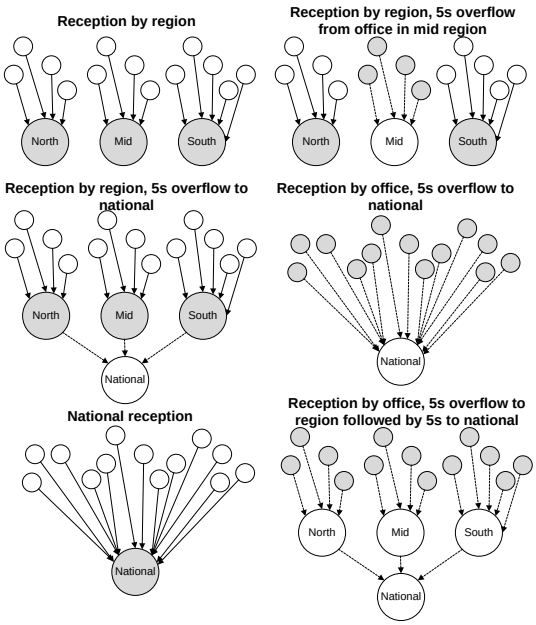


Figure 1.5: Considered traffic routing strategies evaluated in this study. The grey nodes correspond to the initial SBR, and the following nodes are the next pooling stage. The 13 top-circles are the different sites located in Sweden divided into three regions of collaboration, and the bottom circle is a national collaboration queue. The geographical mapping plus the sites that belong to each node is found in Figure 1.4.

research cycle, which briefly consists of exploring, designing, and evaluating. In particular, the exploration aspect seeks to extend the knowledge on call centers as a system and how to manage it. The three objectives are described as follows:

1. To explore the general, service-type unique, and case unique characteristics of inbound call center services
2. To design a simulation model that replicate a call center service sharing the same settings as the Swedish emergency call service, including key characteristics (i.e., skills-based routing and dynamic pooling in a pull system)
3. To evaluate traffic routing strategies in the above-mentioned system setting

1.7 Decomposed and verifiable goals

Objectives 2 and 3 are well-defined and do not require any decomposition. However, those objectives require detailed knowledge about the system. For this reason, Objective 1 is needed, and this objective is more complex since each component of the system needs to be well understood. Some components might require a deeper

investigation for a long period of time in order to be understood. The complexity of determining which component is well understood and has enough detailed information on the system is decided by a validation test of the proposed model. The chosen directions within this research determine which components are investigated. The decision involved three considerations—needing to replicate model behavior, providing a general academic interest, and being manageable within the narrow period of time. Different weight is given to each three consideration, and the decision is reflected in the chosen directions for this study. Within Objective 1, there is an emphasis on achieving a deeper understanding about the arrival process as well as on certain effects that are generally ignored in telecommunication research. These effects can include system modifications that have an indirect effect on system components and stochastic resources. With the above in mind, Objective 1 is divided into 1a to 1c:

1. To explore general, service-type unique and case unique characteristics of inbound call center services
   (a) This thesis expands the commonly applied Poisson process—which assumes stationary intervals—by introducing bursts (see Paper I). Such science is critical in order to understand system behavior.
   (b) Indirect effects from system modifications. The main direction is the service time effect of distancing from its customer (explained in Paper II). There is also a qualitative observation that findings regarding stochastic resources depend on system modifications as well.
   (c) Stochastic resources are needed to replicate the stochastic nature of the system. This is explained further in Paper III.
2. To design a simulation model that replicate a call center service sharing the same settings as the Swedish emergency call service, including key characteristics (i.e., skills-based routing and dynamic pooling in a pull system).
   (a) The designed simulation model is assumed to be good enough if it performs the way it would in real life scenarios with 95% confidence. The verification procedure is in collaboration with SOS Alarm analysts, and the validation is done with statistical inference. These procedures are further explained in Paper III.
3. To evaluate traffic routing strategies in the above-mentioned system setting.
   (a) The evaluation is conducted from an organization perspective, and the evaluation is done with a set of viable traffic routing strategies (as outlined in Paper III.

1.8 Academic contribution

Even though this research was initiated as an applied science project aiming to assist SOS Alarm issues, there are some findings that serve a general interest. For

instance, the thesis presents (a) the importance of emphasizing stochastic agent behavior, (b) negative service time effects of geographically distancing agents from its customer, (c) model latent dependencies among arrivals explaining some of the non-stationarity, and (d) the evaluation of a set of skills-based routing strategies based on geographical location. These are perhaps the findings that have greatest academic interest in this field.

Burst model. The main contribution from this thesis is found in the burst model, which explains some of the non-stationarity of the call arrivals in many call centers. Non-stationarity leads to a naive capacity when planning to meet a certain SLA based on TSF. Burst knowledge is also important when creating supplier contracts with reward and penalties based on TSF. The burst model describes the relationship between call arrival times that report the same event.

Distancing agents from customers. Major trends of call center management include centralization and off-shoring. From a queueing theory perspective, such measures increases efficiency when assuming there are homogenous agents. However, there is a lack of academic evidence on the negative effects. This project show that generally speaking, the further away the agent site is from the customer, the less efficient agent handling is. The result may be interpreted as there are barriers in the form of cultural differences that can grow with the distance, and the cultural distance affects the quality of service. This is particularly important in an emergency call center, where fast and accurate communication are vital.

Routing and pooling strategies. SOS Alarm uses an agent-controlled call allocation feature. This means that the agent decides which calls to answer and when to answer them. This has advantages in an emergency center context, but in general, experts advocate a push functionality. The feature does complicate the implementation of skills-based routing because routing enables call assignment to non-responsive agents. Instead, routing is implemented with an overflow setting, thus giving agents in different classes the ability to answer before being pooled to a more flexible agent class. This type of routing is common in a call center, but it is difficult to estimate the behavior and performance. There is a lack of such assessments in literature as well. This study examines performance among a set of different routing strategies using such a setting.

Stochastic agents. Queueing theoretic models generally assume a stochastic characteristic for arrivals, service time, abandonments, and retrials. The rest of system variables are assumed to be deterministic. This study extends the stochastic drivers by incorporating arbitrary agent behavior. This is essential in skills-based routing using skills groups with a few agents and when performance estimations need to be accurate. Agent behavior is especially important in this case as the time to answer need to be modeled using agent-controlled call assignment.

Chapter 2

Theory

2.1 Call center operations management

Call center management involves a wide range of academic fields. Managers need to plan the production based on an expected load; this can be of interest to certain parties—such as statisticians, queueing theorists, and operations researchers—who build sophisticated models from which decisions derive. In order to make accurate predictions, managers also need to understand their resources and customers. Agent psychology within call centers interests HR academia and Human Interface interests designers and engineers. Managers also need to understand its customer's expectations and how their values are achieved, which adds the economical perspective with its marketing and other consumer psychology aspects.

A hot topic within OR is how to optimize a multi-agent and a multi-customer system—this is referred to as a SBR problem. For instance, this can involve deciding the agent classes that would serve certain customer classes based on some differentiated agent skills and customer need or SLAs. SLAs are quantifiable and are part of KPI; they are typically based on the waiting time of the customer. The most common metrics are ASA and TSF, which is a measure of fraction of calls answered before a predetermined threshold value. Decision models are often based on queueing theory, which includes known logics and mathematical cause-effect relations.

2.1.1 Operational regime

There are some challenges that managers need to consider when designing a call center, and these can include questions such as, *How should a call center be managed? How long of a waiting time is desirable? How long of a waiting time is acceptable? What agent utilization is desirable?* Ultimately, the choice is a trade-off between operational efficiency and service quality. Organizations that use call service have different views regarding its purpose, and thus the call service has different purposes in the organi-

zation's value-creation to its customers. Some organizations uses a call center service as their primary activity, while some only consider it as a supporting activity. The positioning and differentiation of organizations in the market also determine which requirements should be put on the service to maintain the organization's market position. Zeltyn and Mandelbaum (2005) described three different types of philosophical views that a call center operates in: efficiency-driven (ED), quality-driven (QD), and quality-efficiency-driven (QED). The different regimes put different staffing demands and has different performance and efficiency. The regimes are best explained by considering a constant offered load, where:

ED regime would use fever agents with a high utilization and longer waiting times; it is a regime that focuses on efficiency and low operational costs.

QD regime would use more agents with lower utilization but shorter waiting times; it is a costlier regime that focuses more on customer experience.

QED regime uses the strengths of a high agent utilization and short waiting times from both above regimes by increasing the offered load.

The QED regime is desirable for a call center manager, according to Garnett and Mandelbaum (2000). Such a regime is mainly the reason for organizations to centralize and pool their SBR. Pooling is functional when there are several service types or agent skills; instead of using separate queues to each service or skill, the organization uses cross-trained agents and pool all the incoming calls to the same queue. Both pooling and economies of scale can partly explain the emergence of pure call centers, as these centers can use more agent classes and cross-train them to achieve a better efficiency. From a queueing theory perspective, there are undoubted advantages of pooling multi-skilled agents. Its drawbacks comes from the narrow assumptions of queuing theoretic models due to the fact that economies of scale have a relation to HR, as large scales rely on standardizations and control, and such concepts have been proven to affect workers' incentives. Houlihan (2004) argued that "this tension unmasks a series of conflicts: between costs and quality, between flexibility and standardization and between constraining and enabling job design."

2.1.2 Workforce management

After deciding on an appropriate regime, there are operational complex issues that manager face. Mandelbaum and Zeltyn (2006) described the different levels of decisions that call center managers regularly face:

1. Hiring and training existing agents is a proactive action performed with a long time perspective (i.e., resource acquisition).
2. Scheduling the agent classes to meet an expected load is typically a proactive measure, and it can also be reactive as to cope with unexpected changes; this is often performed on a weekly, monthly or few months' time interval (i.e., resource deployment).

3. SBR—which is the act of matching an incoming call to an agent—needs to be dynamically updated per current load, and it is often performed on a daily basis.

These three strategical and operational decision levels affect each other in a downward way. A hiring policy can affect a manager's ability to schedule properly, and the scheduled agents can then decide optimal SBR routines. The hiring policy typically spans over a long time period, and consequently it is based on how many agents are required during each shift. However, the number of agents required on each shift is non-trivial, and this is the focus of most OR in call center. Aside from the three decisions mentioned, there are strategical decisions regarding system design involving agent classes and general SBR routines. They are sometimes affected by the physical location of the agent, which involves economical decisions of call center size and locations. Another strategical issue is designing service protocols and deciding whether they are static or state-dependent. The above-mentioned types of decisions are all non-trivial due to the stochastic nature of the system. Much of OR depends on the queueing-theoretic models, which is a mixture of mathematical modeling and statistics. Generally, inputs in these models are (a) the number of agents, (b) arrival rates, (c) service time, and (d) customer psychology in terms of impatience and abandonments. Mandelbaum and Zeltyn (2006) stressed the impact from human behavior as it also is a non-deterministic factor. Call agents, customers, and production planners all act on some incentives that affect them; these incentives can be affected when making system configurations. For instance, psychological aspects affect the behavior of both operators and customers, which in turn affects stochastic parameters used as input in queueing-theoretic models. All levels of decision making from hiring to scheduling and deciding SBR can affect one another. The interdependency between long and short term planning along with the random elements of the system make operations of a call centers very complex.

2.2 Service time

Mandelbaum and Zeltyn (2006) defined service time as "the time an agent spends handling a call." This includes the actual talk time with the customer but also the potential preparation time and after-call work. A more correct term is average handling time (AHT), which include the time an agent takes to finalize all the necessary parts in a customer contact. AHT is in most cases assumed to be exponential distribution. The exponential distribution holds some characteristics that simplifies calculations in different queueing models, which is explained in section 2.6. However, the exponential assumption has been questioned and challenged; for instance, some studies have shown that the lognormal distribution provides a better approximation (Brown et al., 2005; Bolotin, 2013).

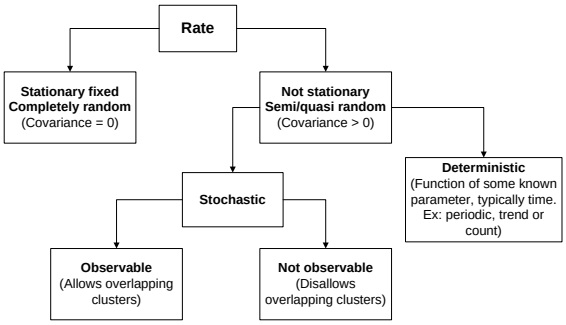


Figure 2.1: Literature review of existing rate models.

2.3 The arrival process

Call arrivals are generally the main stochastic driver to a queue system, and thus they have earned great academic interest during the last decade. Understanding the arrival process is crucial for managers in telecommunication, as arrival characteristics influence capacity requirements. Call center arrivals are generally assumed to be a Poisson process, which have beneficial properties in assisting the later discussed capacity planning. Such properties are that the rate λ is constant (i.e., stationary), and the inter-arrival times are independent and randomly exponentially distributed in nature. Ibrahim et al. (2016) described some properties of assessed call center data that violate the Poisson assumptions; these are namely time-dependence, overdispersion, interday and intraday dependencies, and latent variable dependencies. These violations are managed by treating the arrival rate as a function of time. In the literature, the rate is interpreted according to Figure 2.1.

2.3.1 Time dependence

A main property of call arrivals is that arrival rates generally vary with time; rates are affected by both trends and seasonal variation. Call centers typically have intraday, daily, weekly, monthly and yearly seasonalities (Ibrahim et al., 2016). In addition, there is often a trend that either make a general reduction or increase of calls. As the rate is not constant, the call arrival data is not Poisson in nature. However, this can be accounted for in the non-homogenous Poisson process where the rate is time-dependent, $\lambda(t)$. Such a process is generally applied in call centers, where queueing-theoretic models is solved by using piecewise constant approximations, so that a number of agents are estimated to meet an expected load during an interval. Such an approach is also fundamental in the forecasting procedure. In such an approach, time t is divided into discrete timeslots of length T_{slot} so that there are H intervals

occurring during time T_{tot}, calculated as:

$$H = \frac{T_{tot}}{T_{slot}} \tag{2.1}$$

The expected rate is forecasted from a set of assumed homogenous intervals T_{slot}. The choice of intervals is a trade-off between having short descriptive intervals that distinguish all periodicity while maintaining a high number of data points for accurate expectations. In call centers, these time-slots are generally 15 or 30 minutes in call centers (Aldor-Noiman et al., 2009; Channouf et al., 2007; Avramidis et al., 2004).

2.3.2 Overdispersion

The Poisson process suggests that the variance of arrival counts in homogenous timeslots is the same as the expected number of arrivals during a timeslot. This explains the variability of the Poisson model and can be used as a goodness-of-fit of empirical data to the process. However, there has been observations of larger variances than expectancies (Aldor-Noiman et al., 2009; Channouf et al., 2007). A way to account for the overdispersion is to treat the rate itself as a stochastic variable. Such a model is called a doubly stochastic Poisson process (Ibrahim and L'Ecuyer, 2013; Jongbloed and Koole, 2001; Soyer and Tarimcilar, 2008). In a doubly stochastic Poisson process, the rate $\lambda(t)$ is multiplied with a stochastic variable $\Lambda(t)$ in each timeslot j.

2.3.3 Intra- and interday dependencies

A new problem emerges when $\Lambda(t)$ has time dependent correlation. Gans, Koole and Mandelbaum (2003) believed that much of the rate randomness can explained by covariates between different timeslots. Considering an arrival rate $\lambda_{i,j}$, of day $i = \{1, 2, \ldots, D\}$ and period $j = \{1, 2, \ldots, H\}$, an arrival rate is estimated on H intervals, of length $24/H$ during D days. Studies have shown that there are correlations across both i and j, where j has stronger and longer dependencies (Oreshkin et al., 2016; Channouf and L'Ecuyer, 2012). The reason is related to the underlying reasons of individuals making a call. Within an emergency context, there can be a number of factors that affect the rate; examples include the weather, the fraction of people working, the fraction of people on ski holiday, and other events such as concerts. There are numerous variables that affect an individual's willingness to call the emergency service. The rate is affected by both the probability of an emergency event to occur and the number of people who reacts to the event. These two parameters are affected by an indefinite set of underlying variables which have different characteristics and span differently across i and j. Such a model can imitate periodicity with added randomness as well as day-day and intraday correlations. The phenomenon and a model of such a rate is presented by Oreshkin et al. (2016).

2.3.4 Latent variables dependencies

Arrivals depend on some underlying space consisting of predictive and unpredictable variables that affect the rate of arrivals. Call center managers generally use the predictable variables in their capacity estimation, and ignore the unpredictable ones. This neglect leads to an underestimation of the rate variability. Unpredictable rates can only be assessed probabilistically, and they occur when something happens that triggers several individuals to call. Examples of applications where some arrivals have a dependency are within emergency service (Channouf et al., 2007) and advertisements (Landon et al., 2010).

The unpredictable rates are difficult to assess and understand because the variables are random in both time and magnitude. They are generally not accounted for in capacity planning because they are smoothed out in the forecasting procedure using aggregations. In many application, this may be the most rational approach for a capacity planner, but a model that neglects these effects will likely overestimate the performance.

2.3.5 Forecasting

A reasonable forecasting method needs to account for all above-mentioned issues in order to predict future rates. Expectancies of future rates are generally forecasted from historical call volumes where the forecast model takes seasonality and other factors that affect the rate into account. There are two randomized variables that make this assessment non-trivial: deciding appropriate time segments in which the rate is constant and forecasting the expected number of arrivals within the time segment. The latter is a relatively straightforward task but the former is more complex since the interval length is a random variable where the length of subsequent intervals depend on each other. The fact that interval transitions are often not observable makes the understanding of the phenomenon even more challenging.

Forecasting seasonality is a well-covered field as rate fluctuations are a univariate time-series with several seasonalities. For instance, autoregressive integrated moving average (ARIMA) models and exponential smoothing are recognized methods (Hyndman et al., 2008).). There are also Gaussian models, which are usually built on an autoregressive moving average ARMA that takes the interday and intraday covariates into account (Ibrahim and L'Ecuyer, 2013; Oreshkin et al., 2016). These forecasting methods are used to assess a general base rate for different timeslots. To account for the sensitiveness in the rate calculation and interval length selection, an extension may be added to enhance further variability to the rate in a doubly stochastic approach.

2.4 Abandonments (customer patience)

As seen in the queueing system representation of a call center (see section 1.2), customers who are placed in a queue may leave the queue before being served and will thus not be a burden to the system. This is referred to as impatience or abandonment (Mandelbaum and Zeltyn, 2006). Abandonments are proven to have a large impact on the performance evaluation. This is especially the case in heavily loaded systems where caller have to wait a long time, where abandonments decrease both the load to the system and waiting times compared to the case where all callers would have been served (Garnet et al., 1999; Mandelbaum and Zeltyn, 2006).

Nowadays, the abandonment parameter is often included in queueing models. Generally, the patience of a customer is assumed exponentially distributed so that the individual abandonment rate is ω and the average patience is ω^{-1} (Zohar et al., 2002). There are statistical methods to estimate the patience with different censoring techniques; such methods fall within the concept of survival analysis. The estimation comes from the relationship between the waiting time and the probability to abandon. Mandelbaum and Zeltyn (2006) argued for the importance of including impatience, and they stated that the lack of understanding in these two areas has led to negligence from call center managers. During the last decade, much focus has been on the relationship between expected delay and the probability to abandon (Gans et al., 2003).

2.5 Agent efficiency

A stochastic driver that has received a little attention in literature is the agent efficiency. Individual agents have different work ethic and incentives, and these two highly affect capacity planning. Agent psychology is discussed as an important influencer to OR models (Gans et al., 2003; Aksin et al., 2007). However, it has not yet been explicitly included in decision models; the inclusion has only been made implicit by the manager when assigning schedules and made explicit in sociological HR research. The studies by Batt (Batt, 2002; Batt et al., 2009; Batt and Moynihan, 2002) contain some of the work on traffic management implications from an agent perspective.

2.6 Capacity estimation

A common issue for call center managers is the way to predict the number of agents needed to be scheduled during each time of the day. The issue is with providing the demand of the service with low cost and acceptable waiting times. Gans et al. (2003) defined three different planning levels: (1) queueing models, which determine the number of servers needed to meet accepted waiting times; (2) scheduling models, which determine the shifts of each agent; and (3) hiring models, which determine how many agents are needed in total. These problems need to be emphasized in

increasing order. The first generally is conditioned upon queueing metrics—such as TSF, ASA, and sometimes agent utilization and abandonment rate. The second need is to emphasize other constrictions, such as regulations and an agent's well-being. The third is typically performed with some sort of schedule framework and to some extent arrival periodicity, consequently derived from the second level and accordingly indirect from the first (L'Ecuyer, 2006). Intervals of capacity estimation are generally 30 or 60 minutes in call centers (Gans et al., 2003). The expected number of required agents has a direct dependency to the prediction of the number of arrivals during the given period, and consequently to the forecasting procedure. A common approach in determining the required number of agents for meeting mainly TSF is the Erlang-C formula or Erlang-A. Erlang was an early pioneer in telecommunication; he invented the unit of measure Erlang, which describes the number of trunk lines—also known now as Erlangs—that are needed to cope with the offered load Brockmeyer et al. (1948). The offered load is the average number of calls in the system:

$$E = \frac{\lambda}{\mu} \tag{2.2}$$

where λ is the avarage arrival rate and μ is the average load that each call entails, measured in the same unit. Consequently, $\mu = E[S]^{-1}$ where S is the average service duration. Similarly, the system's average utilization or agent occupancy ρ is derived from the following:

$$\rho = \frac{\lambda}{c\mu} = \frac{E}{c} \tag{2.3}$$

From those metrics, the widely-used Erlang-C formula that expresses the steady-state probability X in which all c agents are occupied is as follows:

$$X(c, E) = \frac{\dfrac{E^c}{c!}}{(1-\rho)\sum_{m=0}^{c-1} \dfrac{E^m}{m!} + \dfrac{E^c}{c!}} \tag{2.4}$$

so that $P[W > 0] = X(c, E)$, where W is the time spent in queue before being served. $P[W > 0]$ indicates the focus of the service with these scenarios: a probability close to 0 indicates a QD regime; a probability close to 1 indicates an ED regime; and a probability strictly in-between 0 and 1 indicates a QED regime (see section 2.1.1).

Assuming inter-day, inter-week and inter-year variations we can derive expected capacity during interval j during day i from a piecewise constant approximation of $\lambda_{i,j}$ and AHT $S_{i,j}$. This is done by calculating the steady-state probability $X_{i,j}$ ($c_{i,j}$,$E_{i,j}$)) in each interval separately. This is a simplified approach that predicts c in an M/M/c queue. Consequently, the formula ignores blocking, abandonments, and retrials. In such a case, the arrival process is assumed to be Poisson with independent exponentially distributed inter-arrival times. The service times are also assumed to be exponentially distributed, and agents c are all statistically-identical. Aside from the Erlang-C formula, Erlang also produced the formula of the probability that a call is blocked in an M/M/c/b queue, with b number of trunks as the Erlang-B formula (Brockmeyer et al., 1948):

$$P_b = \lambda_{E,b} = \frac{\frac{E^b}{b!}}{\sum_{m=0}^{b} \frac{E^m}{m!}} \tag{2.5}$$

A major weakness of the Erlang formulas is the fact that it does not acknowledge abandonments. Palm (1957) further enriched the formula with abandonments by introducing an exponentially distributed patience time with mean θ^{-1} to each arrival. Such a model is defined as a M/M/c+M queue and named Erlang-A, where A stands for abandonments (Mandelbaum and Zeltyn, 2007). It should be noted that the work from Erlang and Palm—that is, the M/M/c and M/M/c+M models—depends on assumptions that are questioned, mainly concerning the arrival process. However, such models are still useful and are commonly applied in capacity planning.

When the manager knows how much capacity is required, a schedule can be created. Scheduling is essentially an optimization problem where agent cost should be at its minimal while meeting the capacity requirement. The simplest integer programming formulation decides whether agent $k = \{1, 2, \ldots, K\}$ is scheduled on shift j on day i so that the total number of agents on shift (i, j) is at least as many as the forecasted capacity $c_{i,j}$. Each agent k has a cost $p_{k,i,j}$ associated to shift (i, j) so that the simplest integer formulation becomes:

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & \sum_{k,i,j} x_{k,i,j} p_{k,i,j} \\ \text{subject to} \quad & \sum_{k} x_{k,i,j} \geq c_{i,j}, \ k = 1, \ldots, K. \end{aligned}$$

There are many more constraints involved, such as constraints that express the longest allowed schedule length and the shortest allowed agent rest. Among others, the work by Avramidis et al. (2010) and Pot et al. (2008) show extensions of the optimization problem.

2.7 Multi-skill and pooling call centers

There are an infinite number of strategical options for the manager when designing the service. Abstractly, there are three types of design methods: decentralize (SBR), centralize (pooling), or a mixture of both. The mixture is based on some condition and is also based on either a load balancing or overflow scheme. This can depend on some specific agent skill or office belonging if the agents are deployed in different places. Gans et al. (2003) referred to such design capabilities as networking, which is facilitated by technology that enables different offices to be managed as a virtual call center. Skills-based routing and networking is enabled by CTI and ACD. The CTI allow telephone and computer to be integrated, thus allowing ACD to function based on call characteristics. In skills-based routing, the ACD is used upon call arrival—that is, it routes the call to an appropriate agent class or in a queue; some examples

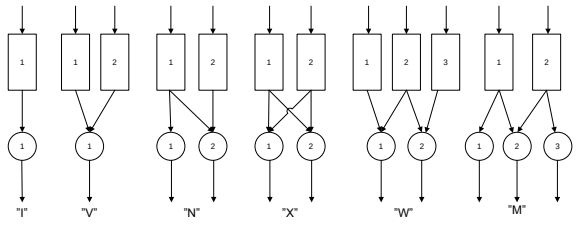


Figure 2.2: Canonical representation of some common designs of skills-based routing (Garnett and Mandelbaum, 2000).

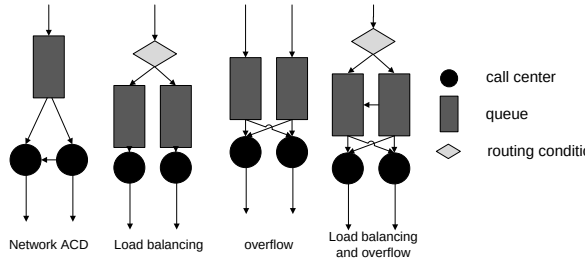


Figure 2.3: The different types and combinations of networking designs (Gans et al., 2003).

of routing are outlined in Figure 2.2. In a network, the call could be routed directly to a queue according to a certain load balancing condition, or the ACD could update the queue after some dynamic or fixed threshold value (an overflow scheme). Figure 2.3 presents some such examples.

2.8 Simulation

Simulation is a common experimental approach where a mathematical model is a representation of the system (Law et al., 2007). Many systems require simulation since they are very complex, and so are the analytical solutions. Such experimental approaches may be necessary when the actual system does not yet exist and when real tests are costly or time consuming. Shannon (1975) stated that simulation is used for understanding the behavior of a system or for evaluating various strategies for the operation of the system. Simulation models are either stochastic or deterministic, static or dynamic, and continuous or discrete (Law et al., 2007). Much of call center evaluation depends on simulation methodology due to the simplified assumptions in analytical theoretic models. The analytical approaches are only useful in simplified queueing settings (Garnett and Mandelbaum, 2000; Gans et al., 2003; L'Ecuyer, 2006). A few of the settings in Figure 2.2 and none of the settings in Figure 2.3 have a closed form analytically. On the other hand, simulation holds capability to in-

clude more variables and can be designed to be as complicated as needed. However, simulation models are dependent on accurate data and random generators that reproduce the data. Within call centers, it is possible to use a generator that reproduces the behavior of service time, call arrivals and abandonments, for instance.

2.8.1 Random number generation

Random variables are mathematically defined as an event that is not predictable, other than probabilistic (L'Ecuyer, 2012). Instead, random numbers are determined by deterministic algorithms using a pseudo-random number generator. Examples of random variables in a call center simulation models could be the time until the next arrival and its service time. Both of these have univariate distributions, which have been well studied in academia. L'Ecuyer (2012) described the process of generating random variates that matches a distribution as a two-step algorithm:

1. Generate variates uniform on the interval zero to one, $U(0\,1)$, which are independent and identically distributed (IID).

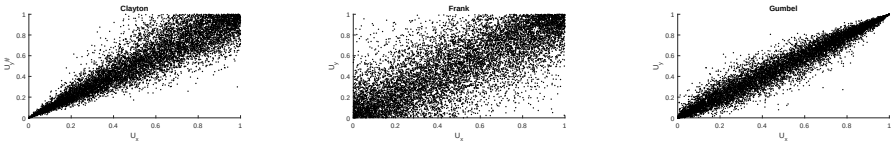
2. Transform the $U(0\,1)$ variates to arbitrary distributions from the inverse of the cumulative distribution function (CDF), known as inverse transform sampling method.

Copula models

Sometimes there is a need to generate $n \geq 2$ variates that have an inter-dependency, and these can be modeled with a copula. Generally these dependents are ignored when generating parameters, as restoring its dependence is a complex task. A copula is a multivariate probability distribution that generates a vector of n random variates with an inherent dependency structure. A copula uses the approximation of two non-decreasing functions transformed to a single dimensional space (Nelsen, 2007). To make this transformation, functions need to be expressed in uniform margins. The CDF is advantageously used due to its non-decreasing property as well as its coherency with the generation algorithm described above. The CDF property is derived from Sklar's theorem (Sklar, 1973).

Examples of copulas are the Archimedean copulas, which enables models of aribtrary dimensions with only one parameter that describes the strength of the dependency structure. Common Archimedean copulas are the Clayton, Frank and Gumbel; they have a percentile relationships as shown in Figure 2.4, where a parameter θ decides the strength of the structure between variable x and y, which is transformed to uniform margins U_x and U_y. The dependence between variables x and y in a Frank copula is mathematically derived from equation 2.6:

$$C_\theta(x, y) = -\frac{1}{\theta} ln(1 + \frac{(e^{-\theta x} - 1)(e^{-\theta y} - 1)}{e^{-\theta} - 1}) \tag{2.6}$$



(a) Clayton copula (b) Frank copula (c) Gumbel copula

Figure 2.4: Example of Archimedean copulas with $n = 10000$ and $\theta = 7$.

When given θ—which is the only parameter that expresses dependence structure—it is possible to generate x conditional on y from equation 2.7:

$$C_x(y) = \frac{\partial C(x, y)}{\partial x} \tag{2.7}$$

2.8.2 Discrete-event simulation

Discrete-event simulation (DES) is a process-oriented approach which is preferable to a model-detailed system interaction when system states alter only at discrete points in time (Law et al., 2007). Such a system model is represented by entities, activity, state, and event. Events alter the state of the system, which then affects entities and activities. A discrete event is advantageously chosen over continuous simulation due to its efficiency by only considering the next event and the current state; only the event times affecting a system state are considered (Law et al., 2007).

2.8.3 Agent-based simulation

Agent-based simulation (ABS) is an agent-oriented approach which is preferable when the system is driven by the entities and their interactions. Entities and agents are intelligent and can make decisions, in contrast to DES where entities are passive and move through a pre-specified system. ABS also handles time as discrete. With intelligent, Law et al. (2007) argues that agents can sense the environment as system states and ant states of other agents, and agents can use the information to make independent decisions. These are modeled by agent attributes and basic if/then rules that condition their behavior. The system's progress is driven by agents' decisions and interactions, in contrast to the more general definition of DES where the system is driven by events. In one way, an agent decision is an event as well. For this reason, many argue that ABS is a special case or variation of DES (Law et al., 2007; Pegden, 2010; Beeker III and Page, 2006; Macal et al., 2013).

Chapter 3

Methodology

3.1 Design science research

The origin of this research project is based on the actual need of an organization that seeks to be more efficient by optimizing its traffic routing routine and by increasing understanding of its system. This illustrates the main purpose of OR, which is to formulate models that assists decision makers using mathematics (Ackoff, 1979). Kruger (2015) questioned the academic contribution of OR since the research incentive is to create a solution towards an individual problem—often at an organization—rather than creating knowledge of objective value. A critic may say that no new knowledge emerge when using existing approaches on new applications. In this sense, the design phase itself cannot be called research, as it only serves to design artefacts for individuals using existing methods and knowledge. However, the process of using knowledge in the design phase and afterwards continues onward to a rigorous evaluation of the artefacts' effectiveness; it is a knowledge-generating process and can thus be called research—or more precisely, design science research (Owen, 1998). Owen (1998) described the process of generating and accumulating

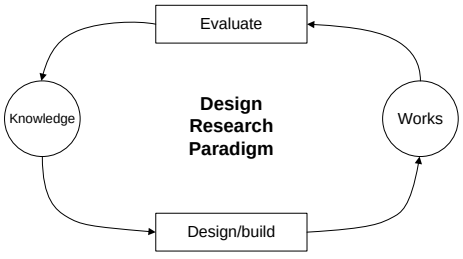


Figure 3.1: DSR cycle according to Owen (1998).

knowledge in a research science paradigm (see Figure 3.1). DSR primarily focuses on the creation of artefacts that assists humans. The underlying driver is not to explain or to understand reality—which is what social and natural scientists often aim to do—but rather to interact with reality using artificial methods (Von Alan et al., 2004). Even though the main objective is to develop a model that assists managers in strategical and operational issues—that is, using DSR—the model requires knowledge about the environment in which the artefact operates, and such a model may include natural and social science contributions. Initially, the objective of the project was to design a model enabling strategical modification of experiments. However, as part of the design process, sub-studies are required in order to investigate service time and the arrival process, both of which are general contributions to science.

3.2 Philosophical standpoint

A number of questions come to mind in the course of the study, such as, is there an objective truth that expresses call center performances? In that case, what is the truth and how is its performance measured? Does the manager who puts up goals and objectives defines values of the service? Or is it the customer or employee? The question may originate from the level of cross disciplinary or perhaps from philosophical theories of research paradigm, truth and reality, and the discussion of ontology, epistemology, axiology and methodology. In the context of OR, an adequate question would be: What is knowledge? Is knowledge something that interests an individual organization? If so, then applied research builds knowledge.

Kuhn defined a research paradigm as "the set of common beliefs and agreements shared between scientists about how problems should be understood and addressed" (Kuhn, 1962). Within OR, these common beliefs and agreements come rather from a pragmatic approach. These are the known factors and factor-dependencies designed in a model that expresses certain quantitative performance metrics. Within OR, the result is considered as the truth and presented to managers as optimums, given that mathematics have a strong objective value. Still, results only rests on the foundation of known factors and relationships. Kruger (2015) stated that "it appears that OR is still mainly inspired by a Newtonian framework that claims that the universe can be understood through a process of reductionism and breaking up of systems into parts in order to understand how the whole system works." The reductionism was challenged by Ackoff (1979) as OR generally does not take into account all the complex interactions from a large number of factors and involved actors.

3.2.1 Research perspective

The degree of objective truth and implications of research findings only applies within the perspective the study takes. Researchers are expected to be aware of the inherent limitations of their used perspectives. For instance, in a question of who defines a system and its values, the discussion can determine how system performance is evaluated and whether it is objective truth or what some believe is the truth. Due

to the methods and purpose that establish the eligibility of OR existence, the community has deeply rooted assumptions regarding call centers, its key performance metrics, and how to evaluate these metrics. Service managers decide organizational directions and metrics and consequently the operationalization is used with a manager's perspective. Manager perspectives are generally customer oriented and HR oriented, and the performance estimation operationalizing the perspectives generally comes from queueing theory.

3.3 Research process

The design science paradigm can be further decomposed into the following operations, which this project as a whole has followed:

1. Conduct an interpretation and formulation of the practical problem
2. Analyze the problem rigorously with theoretical frameworks and research methods
3. Put the findings in relation to current research (as there could be a theoretical contribution)
4. Use theories and findings in the design phase (as there could be a design knowledge contribution)
5. Implement the design into a useful artefact
6. Evaluate the performance of the artefact and its impact

While the findings (Point 3) and design (Point 4) may contribute to the science body of knowledge, the implemented artefact itself do not have any novel contribution. The operations described above is essentially a cycle, similar to the DSR outlined in Figure 3.1. When the last validation is performed, new problems can emerge, and consequently the process starts again at Point 1. It should also be noted that Point 3 is essentially Paper I and Paper II.

At the start of this project, a design research strategy was employed. In order to test system configurations, a mathematical representation of reality allowing experimental was necessary. Designing a model was the main incentive to explore system characteristics, and it required analyses which had not been originally planned. To design a model, the system first needs to be understood, and this requires several deliverables of system entities and actors. Such an assessment requires established methods that explores and describes entities behavior and causalities.

Overall, the methodology follow several steps. To make experimental tests, the study needs a model that replicates the intended system in an appropriate way. The quality of the model is measured by observation and experiments. Simulation is used to test the model's accuracy in order to replicate the performance of the intended system. Simulation is conducted to evaluate the performance before and

Table 3.1: Research overview

Paper	Name	Aim	Applies to	Methodology
I	Modeling bursts in the arrival process to an emergency call center	Explore and design	Stochastic modeling and point processes	Frequentist inference and design
II	Service time effects of distancing from the customer, a case study from the Swedish emergency call center	Explore	Service systems	Frequentist inference
III	Evaluate traffic management strategies at the Swedish Emergency Call Center	Explore, design and evaluate	Call Center managers	Frequentist inference, design and experimental

after routing modifications. When a simulation model that behaves similarly to reality is found, the model is modified, and all model parameters are assumed to be static. This is the critical part, and it is perhaps the main weakness of the design, as it would be impossible to guarantee that all system parameters remain when the system is adjusted.

3.3.1 Paper I

The aim of the study is to initiate a burst model, make initial approximations of the model parameters and consequently propose a methodology of setting up the model. From individual bursts, the study used a maximum likelihood estimation to obtain parameters that fit our burst model. In order to simulate bursts, there needs to be the conditions to allow such parameters with the inherent dependency structure to be generated. The dependence structure was modeled with copulas.

3.3.2 Paper II

The idea to this paper emerged in a variable assessment procedure. There has been hypotheses among organization experts that there are service effects of distancing from the agent from its customer. This study tests the hypotheses and extends them by exploring whether the geographical distance itself is a service quality indicator. The opportunity arose when managers of the service changed to a pooling strategy. The pooling strategy enabled new agent-customer constellations. To compare the performance of agent classes and their relation to customer class, a frequentist inference and hypothesis testing was performed. A pseudo-scale was derived to illustrate whether service time depends on the geographical distance. With the hypothesis, it was shown that with the geographical distance grows the service time.

3.3.3 Paper III

The aim of this paper is to evaluate a set of potential traffic management strategies in order to improve the performance of the service. Due to the complex nature of the routing and pooling setting, there are no analytical forms available, and thus this advocates an experimental approach. Conducting real tests are time-consuming, and some strategies may jeopardize system-critical components and performance, so instead the study designed a DES environment that mimics the behavior of the actual system, and then the strategical modifications can be adjusted in order to compare the different strategies. A drawback of such an approach is the status-quo assumption, assuming all parameters and variables in the model remain constant, except the modification. The study follows the general procedure of simulation modeling (Rossetti, 2015):

1. The problem is formulated and described.

2. The simulation model is designed and finalized.

3. The experimental design's analysis of the modeled system is carried out.

Overall, the intended system is a complex network of queues where detailed information is required about resources, service times, and arrivals. Because of its stochastic properties, the system enters different states depending on previous states. This is the reason the study chose the DES paradigm, which moves dynamically to the next event or state. The study used the commercial simulation tool Arena Simulation Software (Kelton, 2002). The software has useful verification properties together with the supplemented animation features. To guarantee that the model is accurate and reliable, the study conducted qualitative expert observations and quantitative key measurement assessments. The designing phase includes multiple stepwise walkthroughs of the model. Simultaneous output analyses provide average and relative frequencies comparisons to empirical outcome. The uncertainty and acceptance level is based on the central limit theorem. The waiting times of the simulation model was compared to the empirical, and a null hypothesis was formed where the waiting times are equal with a 95% confidence level. Since the waiting time is a descriptive measure, the study stated the hypothesis that:

$\mathbf{H_0}$ The model **is equal** to the system in terms of measure of performance.

$\mathbf{H_1}$ The model **is not equal** to the system in terms of measure of performance.

Sections 4.1 and 4.2 present the model and assumptions therein. Sections 5.1, 5.2 and 5.3 present the modeled variables needed in the simulation model. Section 5.4 presents the validation procedure of testing the hypotheses. Section 5.5 describes the results of objective 3, which is the evaluation of traffic routing strategies. Finally, section 5.6 presents some empirical observations made after strategy modification.

3.4 Data collection

This research project was initiated in collaboration with SOS Alarm Sverige AB, and a part of their commitment was to provide necessary data. SOS Alarm stores every unique call, both answered and unanswered, along with key attributes such as arrival time and calling county. For the answered calls, there are also attributes such as answering office, call classifications, service duration, and a linkage to the parent event if several calls reports the same event. The granularity is a key advantage of this project, where many other services only stores aggregations. However, the level of detail places high demands on data management to maintain integrity and privacy.

3.5 Ethical consideration

Due to the detailed information about emergency calls, the study ensures that no attributes which may be directly or indirectly associated to a civilian are extracted from the system. The mission of the organization and consequently the aim of this research intends to make a positive contribution to society, although it is not clear whether the research would achieve this if an increased efficiency may jeopardize other values. However, this thesis has followed five ethical guidelines for DSR presented by Myers and Venable (2014):

The public interest: Researchers should be aware of all the stakeholders who may be affected by the artefact. This is emphasized by discussing the different perspectives that guide researchers within call center research. This is the foundation of the cross-disciplinary need expressed in literature surveys.

Informed consent: Researchers should obtain an informed consent from anyone who may participate or in some other way interact with the system.

Privacy: Researchers should take all actions to protect the privacy of the individuals involved when interacting with the artefact. All information and every single call is guaranteed to be depersonalized so that no individual can be distinguished in the data.

Honesty and accuracy: The design should not be plagiarized; it should be inspired by other works and should make the research findings publicly accessible.

Property: Researchers should have an agreement of the intellectual property (IP) of the design as well as any findings and information. In this study, researchers must continuously ensure that no private information that may harm individuals or society is disseminated.

Quality of the artefact: All actions should be made to ensure the quality of the artefact. This is especially important in critical systems where all risks should be

mapped and presented. Researchers must also continuously present limitations of each finding so that affected parties actually understand the impact of findings.

Chapter 4

Model

The overall aim of this thesis is to design a simulation model that mimics SOS Alarm's service of inbound emergency calls, in which the system is expressed from a queueing system perspective. This chapter presents the overall model and its general assumptions. In this study, the model is referred to as a DES, and it was modeled in the commercial software Arena Simulation. It is possible that some critics view it as an ABS since the agent behavior that affects the system is also added to the model. However, since ABS is argued to be a special case of DES and the model is designed with DES foundations, the study continues to hold to the viewpoint that the model is a DES.

4.1 Variable assessment

4.1.1 Arrival process

The assessment of a true statistical arrival process that holds all predictive and unpredictable rate transitions is a great undertaking in itself with research that should span several years—even if the model would probably miss some information because such a point process holds periodicity and random rate intervals. There are probably unsolved day-day and intraday correlations as well (see section (5.1). In addition to the above-mentioned reasons, which have already been highlighted, Paper I exploits bursts, which have not been adequately exploited in literature. Since this study did not have the time to solve all these properties, the simulation model uses empirical arrival lists instead of a random generator that holds all of the above mentioned properties. Using an empirical arrival list can hold all the periodicity as well as random rate magnitude and lengths. The drawback is the fact that the result only applies during similar periods—what guarantees that the used call list is representative?

4.1.2 Agent

The system that this study intends to design uses an agent-oriented approach in the call assignment procedure. From a traffic planning perspective, a system-oriented assignment would be easier to understand, plan and control. The reason is that the agent time to answer is a stochastic variable in the system, but it could have been deterministic if the assignment would have been system-oriented. The orientation adds certain values that are important in an emergency service context, such as allowing agents to make situational prioritizations. The drawback is the inclusion of agent behaviors and incentives. For instance, the study modeled the time it takes for an agent to answer when idle with a Beta distribution. It was also necessary to model the two operator states of *responsive* and *unresponsive*. Since there was no empirical data and the system did not support such recording, the study needed to make qualitative assessments. The assessments were validated with empirical performance in order to obtain a reasonable guess (see section 5.3). Regarding queue prioritization, the study assumed FIFO without exceptions.

Another variable that has a high impact on the performance is the capacity (i.e., how many agents are being scheduled). Similar to the arrival process, it is possible to model it as a stochastic variable and assess the expected performance given an uncertain capacity, and this is partly what was done by introducing agent behavior. However, on the whole, the study treats the base capacity as a deterministic variable that is assessed empirically. In this way, empirical arrivals and empirical capacity are used, and only the different routing strategies are varied along with the AHT.

The service time is modeled with parametric distributions and decomposed by call classification, to match the AHT of each service (see section 5.2).

4.1.3 Customer

Another assumption worth mentioning—which can perhaps be a drawback—is the exclusion of abandonments in the simulation model. It can be observed that approximately 10% of the calls are not answered and consequently are abandonments. However, the reason why a customer chooses to abandon is not known. The abandon may be due to impatience and frustration, but it may also be pocket calls and other reasons that not require service. The rationale of abandonment has causation to service time as well as call-back probability. Of emergency call abandonments, it can be noted that a majority are abandoned after just a few seconds. One possibility is that abandonment is due to a lack of patience and typical stress symptoms that may arise in emergency events; another possibility is that it could be a cancelation of unintended calls. The former reason of abandonment is harmful for the organization, while the second prevents unnecessary use of resources. When the system is under stress, an IVR tells the customer not to hang up, as it will result in a longer waiting time if recalling.

The simulation model in this study assumes no abandonments. Empirical arrival lists are used, which only includes the calls answered, and it is assumed that the

same number of calls should be answered by the model with equal performance. The fact that we derive AHT on the answered calls only is something that advocates for exclusion of abandonments. The drawback is that during heavy-load scenarios where the waiting times are high, there would clearly be a relation between waiting time and abandonment. The modeling approach in this study is perhaps the most convenient in an emergency context.

4.2 Schematic description of the model

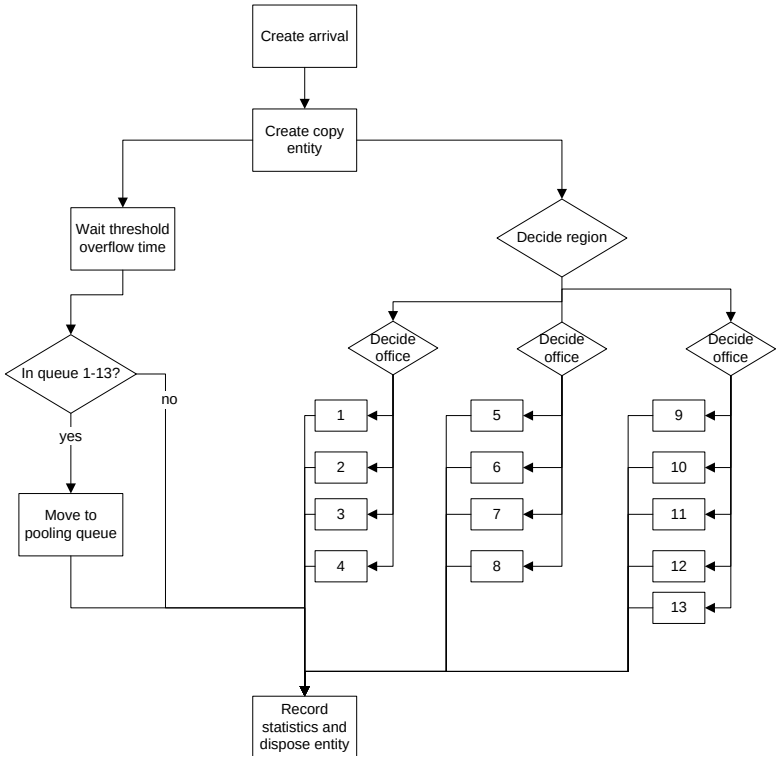


Figure 4.1: Overview of SBR implementation in the simulation model.

The designed model needs to be imitate the SBR settings considered by SOS Alarm system. The routing is deterministically decided upon each customer arrival and based on call type and the geographical location (i.e., determined by the county of the caller position). Offices are assigned calls according to Figure 1.4, and in the simulation model, the call origin is a stochastic variable, and the routing schedule is modeled according to Figure 4.1. Initially there is a region decider, which is needed since some of the evaluated strategies involved clusters of offices, which are called *regions*. The path to the left is a copy entity, which holds all the same information of the call but is put on hold for a threshold value, and eventually, if necessary, it is

moved to the pooling queue. The pooling itself is perhaps the most complex feature within the simulation study. The reason for this is that the study uses conditional queues that are dependent on the waiting-time. With the current implementation, there is a lack of understanding of the effects—that is, the percentage of calls that use the pooling queue and the skills-based queue. The stochastic behavior of the operator also affects the probability of a call to be pooled. Even when a primary-

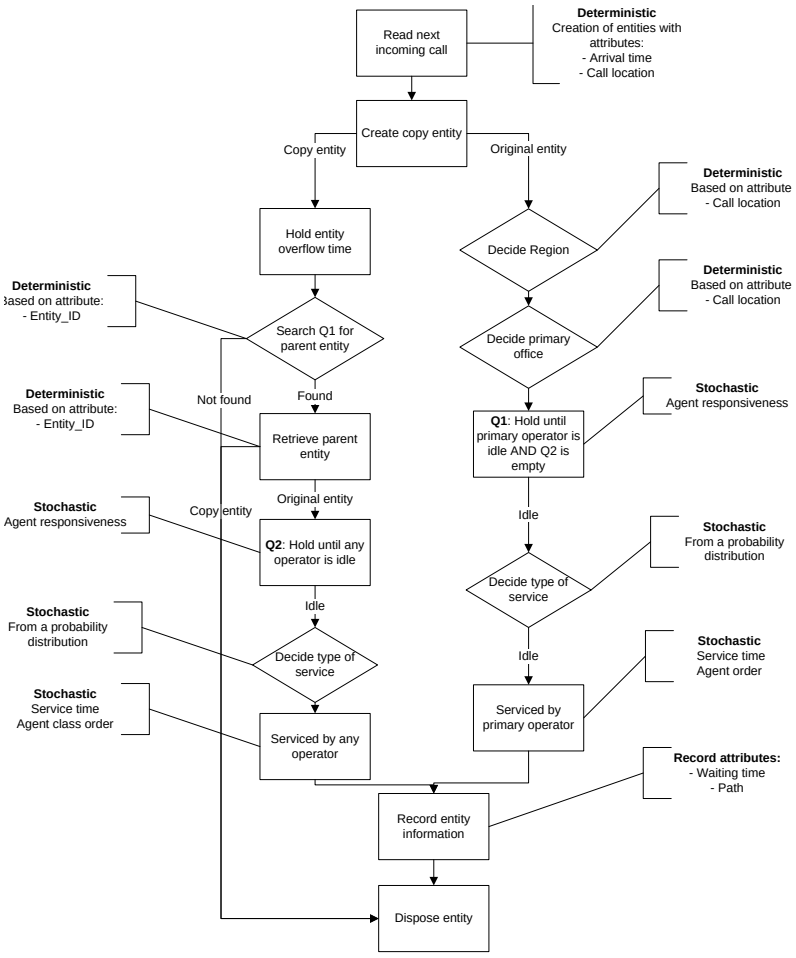


Figure 4.2: Overview of the dynamic overflow/pooling setting implemented in the simulation model, together with stochastic and deterministic assumptions.

skilled agent is idle upon customer arrival, the agent takes a random time to answer the call, which sometimes exceeds the threshold pooling value. However, with an overflow threshold value, a primary-skilled operator would have an advantage to answer before the other agents in the pooling queue because the primary agent has seen the call a longer time. Figure 4.2 shows an overview of queue updates, together with the deterministic and stochastic variables affecting each step.

Chapter 5

Results

5.1 Call arrival process

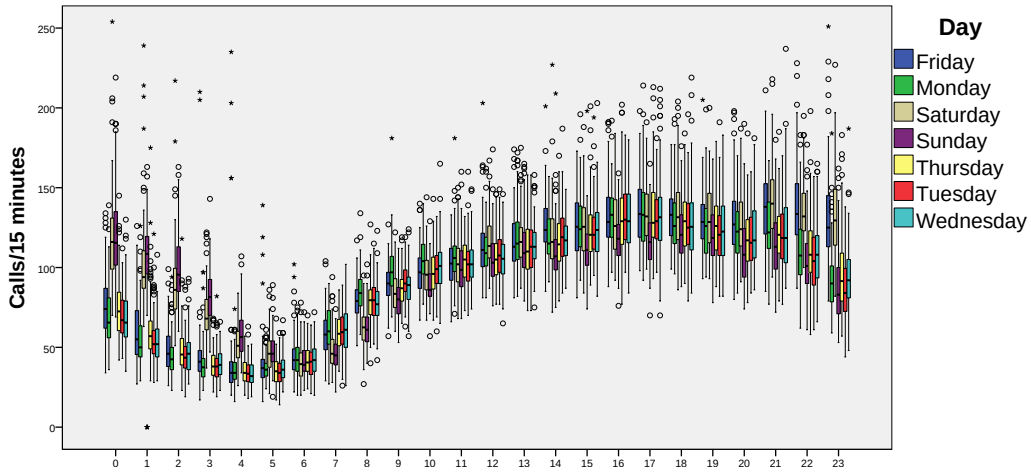
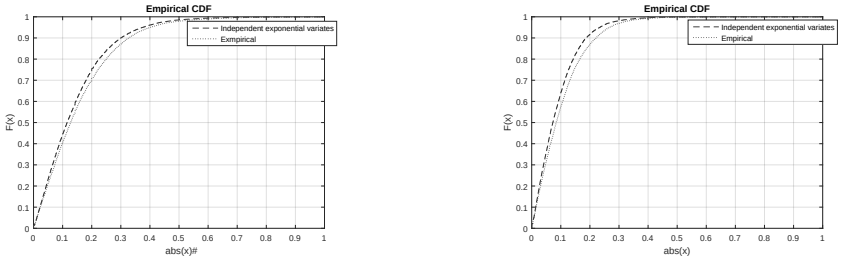


Figure 5.1: Box-plot of arrivals during 15 minute intervals during hours and days. Data from January-June 2016.

As seen in Figure 5.1, the call arrivals follow intraday and interday periodicity, and the difference is distinguished between weekdays and weekends. There is also a certain monthly and yearly periodicity that partly explain the wide boxes in the figure. However, even if it is possible to distinguish periodicity, the rate is still uncertain and random. The wide spread within a homogenous interval is due to other variables affecting the call rate. Some of these can be resolved by modeling a compound Poisson, where the rate itself is multiplied with a stochastic variable. Such



(a) Compared to 5 minute intervals ($k = 6$)

(b) Compared to 10 minute intervals ($k = 3$)

Figure 5.2: Variability within assumed stationary 30 minute intervals, calculated as subset m rate divided by its corresponding interval j rate.

an approach could replicate the overdispersion indicated in Figure 5.2. However, there is another problem not observed in the boxplot, namely non-stationarity or imperfect interval lengths. It is maybe inadequate to determine an interval by a fixed 30-minute timeslot because some of the underlying variables may be active for a shorter interval.

To illustrate the stationarity, the rate variability of $j = \{1, 2, ...n\}$ intervals of length s is compared to a corresponding Poisson process. This is done by splitting interval s into smaller pieces; in this illustration, s is split into k subintervals, and each size is s/k. It is then possible to compare the rate $\lambda_{j,m}$ of interval piece $m = \{1, 2, \ldots, k\}$ in period j to its interval average (λ_j). The total difference is derived from the mean absolute percentage error (MAPE):

$$MAPE = \sum_{j=1}^{n} \sum_{m=1}^{k} \frac{|\lambda_{j,m} - \lambda_j|}{\lambda_{j,m}} \tag{5.1}$$

Figure 5.2 presents the empirical probability of obtaining a rate in a subinterval that differs x in magnitude from λ_j, in comparison to a theoretical model that assumes stationary interval with independent and exponentially distributed inter-arrivals. In the 5-minute interval comparison, the empirical data sets differs on average 15.9% from its half hour average, while exponential variates differ 14.3% on average. In the 10-minute comparison, the empirical differs 10.4% whereas exponential variates differ 8.9%. The CDFs in Figures 5.2(a) and 5.2(b) shows that the tail provides the difference. The tail distribution represents rare events and are a rationale for the work in Paper I.

5.1.1 Burst phenomenon

Figure 5.2 indicates that empirical arrivals incorporates a higher variability than what the Poisson process suggests. It is not possible to explain all of the overdispersion, but it is clearly seen that some come from bursts (see the sudden rate increment in Figure 5.3). Managers can generally assume that the rate within this 2-hour

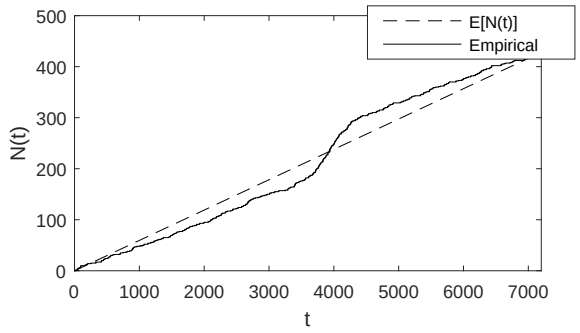
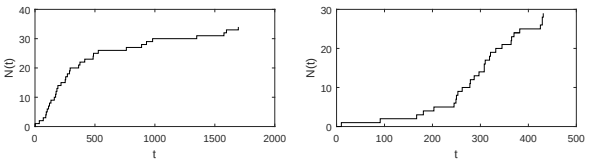


Figure 5.3: Cumulative arrival count of empirical arrivals to the call center compared to expectancy of assuming a stationary period.

period is nearly constant with a small daily variation, but in this example assumes that it is expected to be constant for simplicity. Up to time 3800s, the rate is lower than the manager expectancy, meaning that the service has an overcapacity. Around 400s, the rate is far more than their expectations, which denotes an undercapacity. After 4200s, the rate smooths out to normal, meaning that the service again has an overcapacity. This example shows that managers who assume a stationary interval will likely obtain optimistic SLAs and resource efficiency calculations, which in term leads to unnecessary costs and an inability to achieve SLA. This sudden rate increment was derived from many calls that called for the same event, which can also be referred to as a *burst of calls*.

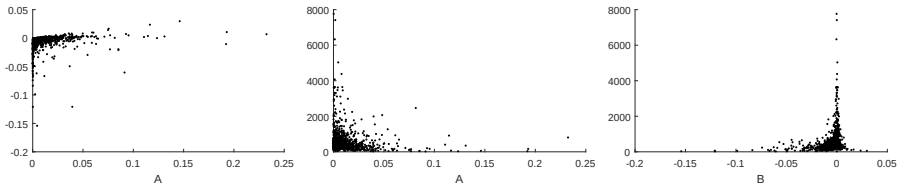
Burst model

Paper I explains how calls linked to a single emergency event are distributed in time after the event. The rate of arrivals at time t after the first call that reports the unique event i is described as $\lambda_{i,t}$. Since the actual time of the underlying event i is not available, t is modeled as the time elapsed since the first knowledge regarding the



(a) Burst with decreasing rate (b) Burst with increasing rate

Figure 5.4: Cumulative count of arrivals that reports the same event.



(a) Relation A, B (b) Relation A, C (c) Relation B, C

Figure 5.5: Empirical dependencies between parameters A, B &C

emergency event i. Two examples of arrivals extracted from a single event is presented in Figure 5.4. Since the rate can either be decreasing, constant or increasing, the rate is modelled as the following:

$$\lambda_{i,t} = \begin{cases} Ae^{-tB}, & \text{if } 0 \leq t \leq C \\ 0, & \text{otherwise} \end{cases} \tag{5.2}$$

where C is the time from first arrival to the last that reports event i (i.e., the length of the burst). A key feature of equation 5.2 is the fact that it represents three observed cases of bursts (see equation 5.3).

$$\begin{cases} B < 0 & \rightarrow \text{increasing rate} \\ B = 0 & \rightarrow \text{constant rate} \\ B > 0 & \rightarrow \text{decreasing rate} \end{cases} \tag{5.3}$$

To incorporate this model into the existing arrival process that describes the occurrence of independent calls i, it is necessary to simulate each burst process independently, meaning that parameters A, B and C of individual bursts are assessed. In Paper I, a maximum likelihood estimation procedure is used to solve each burst separately. The obtained parameters have interdependencies (see Figure 5.5),), and the joint density are modeled with copulas. The complete method of modeling bursts is described in Paper I, and the non-trivial task of estimating the joint density is described below.

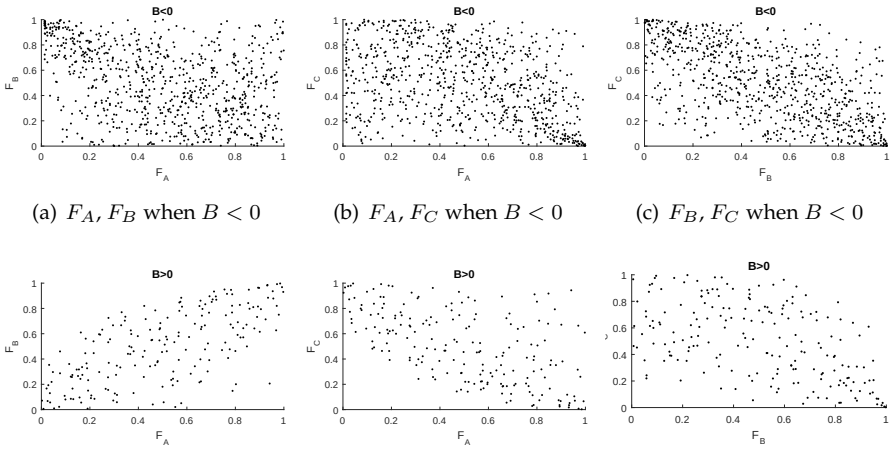
Copula estimation

To model the dependency structure among A, B and C, the paper uses Frank copulas. Frank copulas describe the tail dependencies of each pairs of CDFs. The method of estimating a copula to each combination is described as follows.

Frank copula (see equation 2.6) is fitted to each pair of empirical vectors within $(A_{B<0}, B_{B<0}, C_{B<0})$ and $(A_{B>0}, B_{B>0}, C_{B>0})$:

1. Estimate marginal CDF's (i.e. $F_{A_{B<0}}$, $F_{B_{B<0}}$, $F_{C_{B<0}}, F_{A_{B>0}}$, $F_{B_{B>0}}$ and $F_{C_{B>0}}$) of the estimated parameters based on either the parametric function or a probability density estimate
2. Transform parameters into vectors having uniform margins $U_i = (U_{i,A}, U_{i,B}, U_{i,C}) = (F_A(A_i), F_B(B_i), F_C(A_i))$.
3. Observe the dependencies within U of all bursts, which essentially is the percentile comparison between parameters within a burst (see Figure 5.6)
4. Fit copulas that match the observed dependencies between U_A and U_B, U_A and U_C as well as between U_B and U_C (in this case, Frank copulas with parameters θ_{AB}, θ_{AC} and θ_{BC}, for the two cases $B < 0$ and $B > 0$)

Due to an inverted tail dependency when B goes from negative to positive (see Figure 5.6), separate copulas for $B < 0$ and $B > 0$ are designed.



(a) F_A, F_B when $B < 0$ (b) F_A, F_C when $B < 0$ (c) F_B, F_C when $B < 0$

(d) F_A, F_B when $B > 0$ (e) F_A, F_C when $B > 0$ (f) F_B, F_C when $B > 0$

Figure 5.6: Empirical dependencies of the CDF of parameter A, B and C

Burst model simulation

The above workflow is the method for deciding dependencies between each vector pair (i.e., θ_{AB}, θ_{BC}, and θ_{AC}). In order to generate a set of parameter with the same density, the study first uses $C(A, B)$ to generate A and B with appropriate dependency, and then the same B is used in $C(B, C)$. This is possible since U_A and U_C are rather uniformly distributed, as in A and C are assumed to be independent of each other. Figure 5.6 shows a scatter of all pairs. Such conditional generation is per-

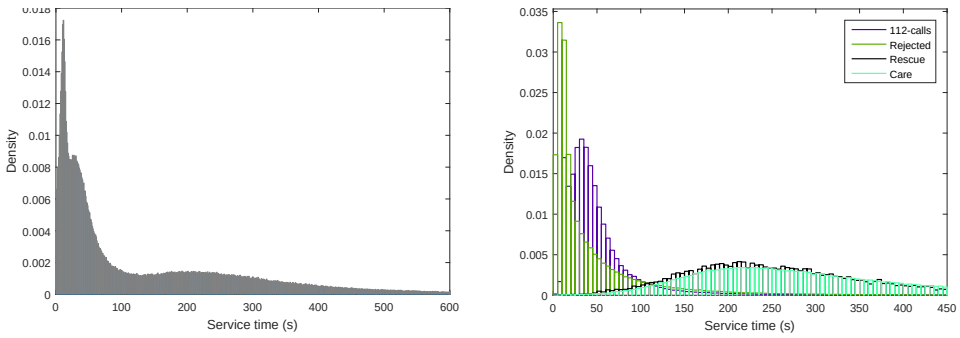
formed with equation 2.7; pairs of (A, B, C) are generated according to the following algorithm:

A two-step algorithm for generating $R = (A, B, C)$ variates from equation 2.7; the algorithm is performed for two cases, namely $B < 0$ and $B > 0$):

1. Draw three independent uniform random variates between 0 and 1 (u_A, v_1, v_2)
2. Set $u_B = -\frac{1}{\theta_{AB}} ln(1 + \frac{v_1(1-e^{-\theta_{AB}})}{v_1(e^{-\theta_{AB}u_A}-1)-e^{-\theta_{AB}u_A}})$
3. Set $u_C = -\frac{1}{\theta_{BC}} ln(1 + \frac{v_2(1-e^{-\theta_{BC}})}{v_2(e^{-\theta_{BC}u_B}-1)-e^{-\theta_{BC}u_B}})$
4. Set $R = (F_A^{-1}(u_A), F_B^{-1}(u_B), F_C^{-1}(u_C))$

After generating a set of parameters with desired multivariate distribution, it is then possible to generate times $T_{a,b}$ of arrival b that belong to burst a as a nestled generation scheme. A nonhomogeneous Poisson process describes the occurrences of a and, with some probability, it starts another nonhomogeneous Poisson process that generates each b from the mathematical model described in equation 5.2 with parameters obtained above.

5.2 Service time

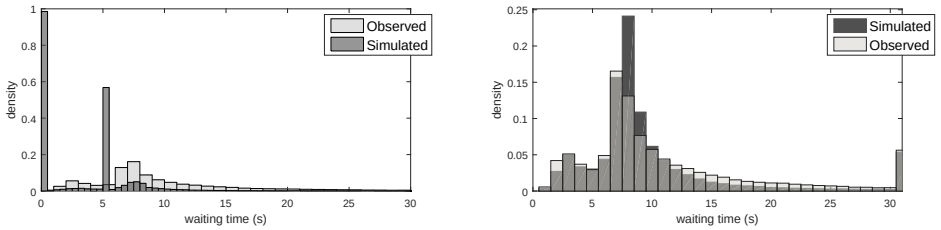


(a) All call classifications

(b) By call classification

Figure 5.7: Empirical service time PDF (2016).

The service time distribution is illustrated in Figure 5.7(a). As seen, it is not a general distribution but rather a compound probability distribution. By distinguishing different call classifications, four exponential distributed functions are obtained. Consequently, the service time follows an exponential distribution where the input variable (i.e., average service time) is a random variable. The random variable is derived from the service time of each independent service classification together with the probability of having that classification. This probability function is time-



(a) Excluding stochastic resources (b) Including stochastic resources

Figure 5.8: PDF comparison between empirical waiting times and corresponding simulation output.

dependent, meaning that there is a larger fraction of calls with a care need during night than day. In other words, the 112-calls are more frequent during days than nights. The four different classifications are distinguished in Figure 5.7(b).

Further insights on the service time is found in Paper II, in which a dependency between the customer and the agent is presented. There are also indications that the service time depends on the distance between the two.

5.3 Agent behavior

The probability density function (PDF) in Figure 5.8 is used as a motivator to incorporate stochastic agent behavior. The empirical waiting times has a larger tail than simulation runs using the same arrivals and resources. Exclusion of agent behavior leads to a more deterministic behavior than what is observed—that is, more calls are answered immediately upon arrival (see Figure 5.8(a)). In the PDF comparison in Figure 5.8(b), the stochastic behavior of the agent is incorporated, as stochastic behavior includes a time-to-answer variable, which on average is 2.5s and best approximated with a Beta distribution. The waiting times were still too much clustered around 2.5s and around 6s to 7.5s . The tail of the waiting time PDF illustrates capacity shortcomings, and thus stochastic absence has been added, according to Table 5.1.

Table 5.1: Operator absence modeled in Arena as failures with up time corresponding to time between absences and down time to the time being absent.

Absence	Operator	Up time	Down time	In state
Recovery and preparation for next call	All	-	NORM(25,10) seconds	All
Longer absence	All	UNIF(1,12) minutes	NORM(60,15) seconds	All
112 operator responsiveness	112-operator	NORM(1.5,0.6) seconds	NORM(14,10) seconds	Idle
Absence due to other duties	Rescue and ambulance coordinator	DISC(0.5,10,1,20) minutes	NORM(10,2) minutes	All
Rescue operator responsiveness	Rescue operator	NORM(1.5,1) seconds	NORM(25,20) seconds	Idle
Ambulance coordinator responsiveness	Ambulance coordinator	NORM(1.5,1) seconds	NORM(25,20) seconds	Idle

5.4 Simulation model validity

To validate the model proposed in section 4 along with the estimated variables, the study performs a hypothesis test to check whether the empirical and simulated waiting times are equal. It should be noted that subsequent waiting times are not independent of each other. If a customer waits a long time, it probably means that all agents are occupied, and reasonably the following arrival has a great probability to wait a long time as well. In short, there is a higher probability of waiting a long time during periods when the system is strained than others. From the same logic, the waiting times are not distributed equally during periods with idle agents and during periods with long waiting times. Those two properties are important assumptions of statistical inference and are called IID random variables. These assumptions need to be fulfilled in simulation model validation where a comparison of the models output is compared to corresponding empirical observation. To convert the waiting times into an IID process, the following steps are performed:

1. Log-transform the empirical waiting times as well as corresponding model output, which are the waiting times from the simulation model.

2. Divide both processes into 100 batches with 1119 points in each

The logarithmic data set approaches a normal distribution, as the standard deviation includes $\approx 70\%$ of the sample compared to absolute waiting times—including $\approx 90\%$ of the sample—in contrast to the desired $\approx 68\%$, which applies to a normal distribution. The mean difference of the logaritmic empirical and simulated batches is calculated to be -0.0413. The standard deviation of logarithmic pair-differences is 0.025. The standard deviation stipulates that the value of z is approximately 1.65. Essentially, around 1.65 standard deviations is needed to cover $\mu = 0$, and since the two-tailed test is of main interest here, it can be stated that the model cannot be rejected with a confidence level of 95%.

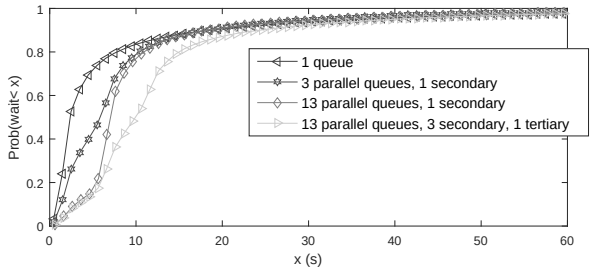


Figure 5.9: CDF comparison of the evaluated routing strategies.

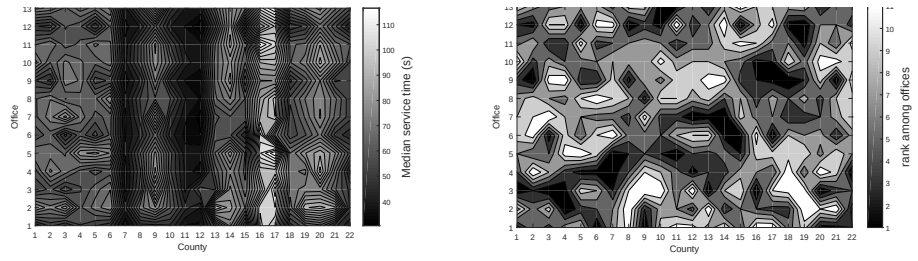
Table 5.2: Full table of routing strategy performance based upon 15 simulation replications, n=33318.

	Performance						Routing effects							
	Wait time (s)		[%] within 30s		Max wait (s)		[%] primary office		[%] primary region		[%] multi-skill		[%] overflow queue	
Strategy (reception according to)	A	SD	A	SD	A	SD	A	SD	A	SD	A	SD	A	SD
National	10.0	1.4	93.8%	0.6%	461.4	203.0	-	-	-	-	4.8%	0.3%	95.2%	0.3%
Regional to national	11.2	0.7	93.4%	0.6%	290.1	63.8	-	-	49.5%	0.5%	3.2%	0.1%	47.3%	0.5%
Regional 1	12.2	1.5	93.9%	0.4%	641.2	157.2	-	-	100.0%	0.0%	-	-	-	-
Office to national	12.4	0.7	93.8%	0.9%	315.7	117.7	19.1%	0.3%	-	-	3.4%	0.1%	77.5%	0.3%
Regional 2	13.3	2.3	93.1%	0.5%	599.5	187.0	-	-	100.0%	0.0%	-	-	-	-
Regional except middle offices	13.6	1.6	93.1%	0.5%	533.4	103.6	13.5%	0.3%	86.5%	0.3%	-	-	-	
Office to regional to national	15.0	0.8	92.1%	0.7%	319.0	98.7	18.4%	0.2%	38.6%	0.4%	3.0%	0.1%	40.0%	0.5%
Office to national	6.3	0.0	99.9%	0.0%	51.2	11.4	41.8%	0.3%	-	-	0.2%	0.1%	57.9%	0.3%

Regional 1 = 150% staffing, Regional 2 =120% staffing and modified multi-skill contribution.
A = average of 15 replication runs, SD = standard deviation

5.5 Routing and pooling effects

From a queuing-theoretic perspective, the positive effects of pooling is undeniable. The more agent classes collaborate (i.e., the larger pooling class), the more efficient it would be. Therefore, the more centralization there is, the higher the probability of a lower ASA and a better TSF would be (see Figure 5.9). An important point to note is that the term *efficient* is based on ASA and TSF. For this, there is a theoretical basis, but it derives from a queuing theoretic perspective. The term *quality* denotes more than merely waiting time. For instance, the service time may affect the customer experience. Aside from this, the service time may be affected by exploiting less skilled agents. Paper II show that there is a statistical difference of service time between agents' primary county and the same agents taking calls from outside their primary county. However, this effect is not observed among all operator classes; only five out of 13 offices show a significant improved service time in primary county assessment. By contrast, five out of 13 agent classes have longer service times in their primary counties than outside their counties. The remaining three are not statistically significant. There are ongoing discussions on the reason of the alteration and whether



(a) Service time

(b) Agent ranks of each customer class, adjusted by customer class average

Figure 5.10: Empirical service time of county (customer class) and site (agent class)

they are avoidable by further technical support and training. The median service has structural differences (due to office-specific or county-specific characteristics) as seen in Figure 5.10(a). This should be modeled in the simulation model, but such knowledge did not exist prior to the design decision. Figure 5.10(b) illustrates a potentially distance-dependent difference by showing the rank of counties by each agent class. In the rank comparison, the structural differences have been adjusted for (see Paper II). The performance of each strategy is outlined in Table 5.2.

Chapter 6

Discussion

6.1 Routing effects observed empirically

Managers at the call center have used the designed simulation model and the results from traffic management strategy validation to make routing modifications at the end of 2015. They implemented the most decentralized skills-based routing combined with a pooling setting, allowing a call to be answered by all agents after five seconds. The actual test allowed validation of the simulation model assumptions. Initially, the effects were unequivocally positive and corresponded to the simulation results. TSF improved by around 10% to 95-99%, and the agent class workload got a more even spread. A year after the strategic modification in early 2017, the distribution of calls and agents were according to Table 6.1.

After a few months of the national pooling strategy, the performance started to decline, and eventually it decelerated at the same level as prior to the strategical shift. There are plenty of variables that are not controllable, which makes any attempt to derive the effect impossible. For instance, there could be a different workload, capacity, new agents, or new working protocols. However, at least when adjusting for capacity and workload, the declining trend was still observable. Workshops with organization specialists suggested that there is probably a change in agent incentives and attitude. This is supported by HR research that suggest that small teams provide a greater sense of belonging and responsibility. A reason is that the strategy implementation implies that the HR strategy moves from involvement and commitment to control-oriented, which is argued to have negative effects on agents (Batt, 2002; Batt and Moynihan, 2002).

Another hypotheses is that the service quality is jeopardized as less skilled agents (from another site) are utilized. In queueing theory, the term service quality is most often described by the service time, which perhaps is suitable in an emergency context where customers would prefer to have quick guidance and emergency units sent to the correct location as soon as possible. At SOS Alarm, the importance of correct and professional service extends quality to other characteristics, such as correct

Table 6.1: Empirical call allocation during September 2016 to August 2017

	Falun	Goteborg	Halmstad	Jonkoping	Karlstad	Lulea	Malmo	Norrkoping	Stockholm	Sundsvall	Vaxjo	Orebro	Ostersund	total nr
Blekinge county	5%	9%	6%	6%	4%	3%	8%	9%	8%	2%	31%	3%	6%	47 921
Dalarnas county	33%	8%	5%	6%	4%	3%	7%	8%	8%	3%	5%	3%	9%	91 850
Gotlands county	6%	10%	7%	7%	5%	3%	9%	10%	27%	2%	6%	3%	7%	15 686
Gavleborgs county	33%	8%	5%	5%	4%	3%	7%	8%	8%	3%	5%	2%	9%	90 532
Hallands county	5%	9%	28%	6%	4%	3%	9%	9%	9%	2%	6%	3%	6%	89 155
Jamtlands county	4%	8%	5%	5%	4%	3%	7%	7%	7%	3%	4%	2%	41%	38 382
Jonkopings county	5%	9%	7%	29%	5%	3%	9%	9%	9%	2%	6%	3%	6%	103 625
Kalmar county	5%	9%	6%	6%	4%	2%	8%	9%	9%	2%	30%	3%	6%	79 479
Kronobergs county	5%	9%	7%	6%	4%	2%	8%	9%	8%	2%	30%	3%	6%	60 384
Norrbottens county	4%	6%	4%	4%	3%	49%	6%	6%	6%	2%	4%	2%	6%	79 111
Skane county	5%	9%	7%	7%	5%	3%	29%	9%	9%	2%	6%	3%	6%	432 594
Stockholms county	5%	9%	7%	7%	5%	3%	9%	10%	28%	2%	6%	3%	6%	783 631
Sodermanlands county	5%	9%	7%	7%	5%	3%	8%	30%	9%	2%	6%	3%	6%	107 646
Uppsala county	34%	8%	6%	5%	4%	3%	7%	8%	8%	3%	5%	2%	9%	91 543
Varmlands county	3%	6%	4%	4%	56%	2%	5%	6%	5%	1%	3%	2%	4%	86 849
Vasterbottens county	4%	8%	5%	5%	4%	2%	7%	7%	7%	3%	4%	2%	41%	73 641
Vasternorrlands county	4%	8%	5%	5%	4%	3%	7%	8%	8%	32%	5%	3%	8%	80 932
Vastmanlands county	5%	9%	6%	6%	5%	2%	8%	9%	8%	2%	5%	31%	6%	91 377
Vastra Gotalands county	5%	29%	7%	7%	5%	3%	9%	9%	9%	2%	6%	3%	6%	557 675
Orebro county	5%	8%	6%	6%	4%	2%	8%	8%	8%	2%	5%	32%	6%	99 075
Ostergotlands county	5%	9%	7%	7%	5%	3%	8%	29%	9%	2%	6%	3%	6%	147 475
other	11%	9%	6%	6%	4%	5%	8%	8%	9%	4%	5%	3%	22%	62 796
total nr	251 956	402 778	226 898	229 270	194 790	124 413	362 284	346 701	441 975	94 226	224 635	149 273	260 933	3 311 359

positioning and quality assurance regarding service process and emergency assessment. There were also hypotheses that place-specific knowledge and perhaps language barriers prevent efficient agent handling and jeopardized the service quality. In Paper II, this is analyzed by assessing the service times of calls where primary and secondary agents were used. The results show some differences but not a major one that would explain the declining performance. There are probably more factors that did not correspond to this study's assumptions of status quo. Variables and parameters vary over time, and many of them depend on the system configuration.

6.2 Research limitations

At the start of this research and probably in DSR generally, it was not clear which directions would be taken in this study or whether any findings would be of value to a broader audience. It was understood that many findings would most likely only be valuable to the studied call center. At the same time, the study aimed to identify knowledge that could interest a wider academic audience. The implication of each finding is decided from the novelty and applicability of that particular finding, which may only be uncovered when far into the procedure, as is typical in a design research methodological framework. This may be the reason why applied science and OR have been challenged as a research discipline by some, as Kruger (2015) argued. What constitutes novel research is a debated question that is difficult to answer. However, what limits this research is easier to discuss, and in the case of this study, the limitations lie in the research perspective, reductionism, and status quo assumptions.

Reductionism and status quo

Other limitations of this research are derived from the reductionism issue described by Ackoff (1979). By reductionism, the system is broken into smaller components—often assumed to be independent—and prior knowledge about similar components

is applied. These simplifications are required due to the complex nature of the systems and a lack of information and time to control all variables. Reductionism has some conformability issues, namely (a) an applicability to apply general assumptions on unique system component, (b) the risk to overlook critical system components, and (c) overlooking interdependencies among system components. Status quo assumption refers to the fact that only knowledge and causations known at the design of the model are taken into consideration. There are several factors and effects that are yet to be discovered. Therefore, it would be more constructive to argue for the simulation model to be optimized rather than the system because the optimum is based on the simulation model assumptions and the knowledge put into it. Overall, it is certain that the assumptions made in the simulation model have weaknesses. For instance, dependence between traffic routing strategy and agent behavior. To overcome such biases, the call center management needs to understand both agent and customer psychology, which is most likely impossible to quantify prior a decision implementation.

Research perspective

An ontology and epistemology discussion was mentioned earlier regarding whether there is an objective truth about quality measurements of call centers. In general, no objective truths exists in this regard because there are always different perspectives. Managers, agents, and customers have different values, and all these contribute to a defined performance metrics. The importance of each stakeholder value is decided by managers themselves as they state the objectives. How the translation is performed and how much each sub-goal will contribute will likely differ depending on the service and the profile of the company. A profit-making company should emphasize their profile and make a trade-off between service quality, waiting time, and operational costs. Monetary terms generally have the overall impact in OR as it is the rationale for an organization's existence. On the other hand, a public service may value the benefit of its citizen higher than other values. In an emergency call center, this is further divided into valuation skillful agents and consequently low agent turnover, which puts more effort on an agent's well-being than perhaps private companies. The problem is that queueing theoretic models do not take HR nor customer perspectives into account more than the expected utilization of agents—and in turn customers waiting and service times—in a rather tayloristic thinking. Without doubt, managerial decisions affect the emotional states of customer and agents, which for instance might affect customers' perceived service and agent quit rate.

6.3 Future work

6.3.1 Cross-disciplinary approach

This research stresses the importance of extending the system boundaries in call center OR. In order to obtain reliable long-term results and effects, a bigger map of

dependent and independent variables is needed. This includes the mapping of variables from HR and CRM since there are interdependencies spanning across these research disciplines. In current queueing theoretic models, strategies and schedules are seen as independently controllable variables, and the expected performance derives from statistical relationships between arrivals, service time, waiting time, and abandonments. On the other hand, HR research urges that strategies and scheduling affect resources (see Batt and Moynihan (2002); Batt (2002); Batt et al. (2009). Furthermore, resources have also been proven to affect the performance, which also this study has proven. Consequently, the queueing theoretic models of call centers uses a narrowed and simplified view. Narrow assumptions may provide incorrect optimum or they may be missing other important effects such as staff turnover. It is clear that much of such dependencies concern HR and marketing research, and this line advocates a multi-criteria decision analysis.

6.3.2 Implications of exploiting bursts

Paper I explained the concept of bursts and a method to parametrize bursts. The significance and implication of such models is however yet to be found. There is a need to include such a model in current arrival models in order to see how much of the variability can be explained by bursts as well as how it improves performance estimation. This is of interest to call center managers but also to the mathematical field of point processes.

From an operational manager perspective, it would be interesting to combine the burst model with situational strategies. This may involve invoking IVR or a modified routing feature in case of a burst. Such a method includes detection of bursts in real-time, which might be a difficult task. However, if such triggers are found, implications of situational routines can be statistically derived, which may be of great interest to call center managers.

Chapter 7

Conclusions

7.1 Research findings

This thesis presents research that aims to assist the Swedish Emergency call center, SOS Alarm Sverige AB. The overall aim is to provide insights and decision support to managers of the service. The aim required an exploration of system variables, a design of a simulation model, and an evaluation of different traffic management strategies, which is done in the experimental environment. In both the exploration and evaluation phase, there are some key findings with academic interest. A main contribution comes from the extension to model stochastic call arrivals by exploiting underlying emergency events. Other contributions are the findings of certain important causalities between traffic management strategy, service quality, and agent behavior.

Burst model

As the author, I would argue that the main contribution from this research project is found in the burst model, which explains some reasons of the non-stationarity of the call arrival process that complicates performance estimations in many call centers. Non-stationarity causes an under-dimensioning of the capacity by managers when planning to meet a certain SLA based on TSF. Burst knowledge is also important when creating supplier contracts with reward and penalties based on TSF. The burst model describes the relationship between call arrival times that report the same event. Such a model is unique in call center research. This thesis and Paper I in particular describe a mathematical model that explains these bursts; it is a methodology on how model parameters can be estimated and a methodology to generate parameter variates.

Distancing agents from customers

Major trends of call center management are centralization and off-shoring. From a queueing theoretic perspective, such measures increases efficiency. However, there is a lack of academic evidence on negative effects. This study shows that generally speaking, the further away the agent is from the customer, the greater the service time is, which can be interpreted as a service quality violation. The result may be interpreted as there are barriers in form of cultural differences that grow with the distance, and the cultural distance affects the quality of service. This is particularly important in an emergency call center, where fast and accurate communication is vital.

Stochastic agents

Queueing theoretic models generally assume that performance of call centers can derive from stochastic arrivals, service time, abandonments, and retrials. The rest of the system variables are assumed to be deterministic. This study extends the stochastic drivers by incorporating arbitrary agent behavior. This is essential when using overflow pooling in combination with agent-controlled call assignment because agent control leads to a random delay, which sometimes leads to pooling while having an agent idle. Stochastic resources are also important when performance estimations need to be accurate since delays affect SLAs. For instance, this study's model needed to incorporate agent behavior in order to obtain statistical significant equally waiting times, which proves its significance.

Pooling strategies

SOS Alarm uses an agent-controlled call allocation—that is, the agent decides which calls to answer and when to answer them. This has advantages in an emergency center context, but in general, experts advocates a push functionality. The agent-oriented call assignment complicates the implementation of skills-based routing, because routing enables call assignment to non-responsive agents. Instead, routing is implemented with an overflow setting, giving agents who are higher up in the hierarchy the opportunity to answer before the call is pooled to a more flexible agent class. This type of routing is common in call centers, according to o Gans et al. (2003), even though there is a lack of empirical evidence regarding its performance. However, this study compares the performance among a set of different routing strategies using such a setting, thereby resolving a part of the complex dynamics of routing and overflow pooling.

Bibliography

Ackoff, R. L. (1979). The future of operational research is past. *Journal of the operational research society*, pages 93–104.

Aksin, Z., Armony, M., and Mehrotra, V. (2007). The modern call center: A multi-disciplinary perspective on operations management research. *Production and operations management*, 16(6):665–688.

Aldor-Noiman, S., Feigin, P. D., and Mandelbaum, A. (2009). Workload forecasting for a call center: Methodology and a case study. *The Annals of Applied Statistics*, pages 1403–1447.

Avramidis, A. N., Chan, W., Gendreau, M., L'Ecuyer, P., and Pisacane, O. (2010). Optimizing daily agent scheduling in a multiskill call center. *European Journal of Operational Research*, 200(3):822–832.

Avramidis, A. N., Deslauriers, A., and L'Ecuyer, P. (2004). Modeling daily arrivals to a telephone call center. *Management Science*, 50(7):896–908.

Batt, R. (2002). Managing customer services: Human resource practices, quit rates, and sales growth. *Academy of management Journal*, 45(3):587–597.

Batt, R., Holman, D., and Holtgrewe, U. (2009). The globalization of service work: Comparative institutional perspectives on call centers: Introduction to a special issue of the industrial & labor relations review. *ILR Review*, 62(4):453–488.

Batt, R. and Moynihan, L. (2002). The viability of alternative call centre production models. *Human Resource Management Journal*, 12(4):14–34.

Beeker III, E. R. and Page, E. H. (2006). A case study of the development and use of a mana-based federation for studying us border operations. In *Proceedings of the 38th conference on Winter simulation*, pages 841–847. Winter Simulation Conference.

Bolotin, V. (2013). Telephone circuit holding time distributions. In *Proc. 14th Int. Conf. Fundamanetal Role of Teletraffic in the Evolution of Telecommunications Networks*, volume 1, pages 125–134.

Brockmeyer, E., Halstrm, H., Jensen, A., and Erlang, A. K. (1948). The life and works of ak erlang.

Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S., and Zhao, L. (2005). Statistical analysis of a telephone call center: A queueing-science perspective. *Journal of the American statistical association*, 100(469):36–50.

Channouf, N. and L'Ecuyer, P. (2012). A normal copula model for the arrival process in a call center. *International Transactions in Operational Research*, 19(6):771–787.

Channouf, N., L'Ecuyer, P., Ingolfsson, A., and Avramidis, A. (2007). The application of forecasting techniques to modeling emergency medical system calls in calgary, alberta. *Health care management science*, 10(1):25–45.

Cinlar, E. (2013). *Introduction to stochastic processes*. Courier Corporation.

Feinberg, R. A., Hokama, L., Kadam, R., and Kim, I. (2002). Operational determinants of caller satisfaction in the banking/financial services call center. *International Journal of Bank Marketing*, 20(4):174–180.

Gans, N., Koole, G., and Mandelbaum, A. (2003). Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management*, 5(2):79–141.

Garnet, O., Mandelbaum, M., and Reiman, M. (1999). Designing a telephone call center with impatient customers. *Submitted for publication in MSOM*.

Garnett, O. and Mandelbaum, A. (2000). An introduction to skills-based routing and its operational complexities. *Teaching notes*, 114.

Houlihan, M. (2004). *Tensions and variations in call centre management strategies*. Springer.

Hyndman, R., Koehler, A. B., Ord, J. K., and Snyder, R. D. (2008). *Forecasting with exponential smoothing: the state space approach*. Springer Science & Business Media.

Ibrahim, R. and L'Ecuyer, P. (2013). Forecasting call center arrivals: Fixed-effects, mixed-effects, and bivariate models. *Manufacturing & Service Operations Management*, 15(1):72–85.

Ibrahim, R., Ye, H., L'Ecuyer, P., and Shen, H. (2016). Modeling and forecasting call center arrivals: A literature survey and a case study. *International Journal of Forecasting*, 32(3):865–874.

Jongbloed, G. and Koole, G. (2001). Managing uncertainty in call centres using poisson mixtures. *Applied Stochastic Models in Business and Industry*, 17(4):307–318.

Kelton, W. D. (2002). *Simulation with ARENA*. McGraw-hill.

Kendall, D. G. (1953). Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded markov chain. *The Annals of Mathematical Statistics*, pages 338–354.

Koole, G. and Pot, A. (2006). An overview of routing and staffing algorithms in multi-skill customer contact centers. *Submitted for publication*.

Kruger, H. A. (2015). *The ontology of Operations Research and Complexity Theory: a critical analysis*. PhD thesis, North-West University (South Africa), Potchefstroom Campus.

Kuhn, T. S. (1962). 1970. *The structure of scientific revolutions*, pages 31–65.

Landon, J., Ruggeri, F., Soyer, R., and Tarimcilar, M. M. (2010). Modeling latent sources in call center arrival data. *European Journal of Operational Research*, 204(3):597–603.

Law, A. M., Kelton, W. D., and Kelton, W. D. (2007). *Simulation modeling and analysis*, volume 3. McGraw-Hill New York.

L'Ecuyer, P. (2006). Modeling and optimization problems in contact centers. In *Quantitative Evaluation of Systems, 2006. QEST 2006. Third International Conference on*, pages 145–156. IEEE.

L'Ecuyer, P. (2012). Random number generation. In *Handbook of Computational Statistics*, pages 35–71. Springer.

Macal, C. M., North, M. J., and Samuelson, D. A. (2013). Agent-based simulation. In *Encyclopedia of Operations Research and Management Science*, pages 8–16. Springer.

Maglio, P. P. and Spohrer, J. (2008). Fundamentals of service science. *Journal of the academy of marketing science*, 36(1):18–20.

Mandelbaum, A. and Zeltyn, S. (2006). Service-engineering of call centers: Research, teaching, practice. In *IBM Business Optimization and Operations Research Workshop, Haifa, Israel*.

Mandelbaum, A. and Zeltyn, S. (2007). Service engineering in action: the palm/erlang-a queue, with applications to call centers. *Advances in services innovations*, pages 17–45.

Mehrotra, V. (1997). Ringing up big business. *OR MS TODAY*, 24:18–25.

Myers, M. D. and Venable, J. R. (2014). A set of ethical principles for design science research in information systems. *Information & Management*, 51(6):801–809.

Nelsen, R. B. (2007). *An introduction to copulas*. Springer Science & Business Media.

Oreshkin, B. N., Réegnard, N., and L'Ecuyer, P. (2016). Rate-based daily arrival process models with application to call centers. *Operations Research*, 64(2):510–527.

Owen, C. L. (1998). Design research: Building the knowledge base. *Design Studies*, 19(1):9–20.

Palm, R. C. A. (1957). *Research on telephone traffic carried by full availability groups*. Tele.

Pegden, C. D. (2010). Advanced tutorial: overview of simulation world views. In *Simulation Conference (WSC), Proceedings of the 2010 Winter*, pages 210–215. IEEE.

Pot, A., Bhulai, S., and Koole, G. (2008). A simple staffing method for multiskill call centers. *Manufacturing & Service Operations Management*, 10(3):421–428.

Rossetti, M. D. (2015). *Simulation modeling and Arena*. John Wiley & Sons.

Saaty, T. L. (1961). *Elements of queueing theory: with applications*, volume 34203. McGraw-Hill New York.

Shannon, R. E. (1975). Systems simulation; the art and science. Technical report.

Sklar, A. (1973). Random variables, joint distribution functions, and copulas. *Kybernetika*, 9(6):449–460.

Soyer, R. and Tarimcilar, M. M. (2008). Modeling and analysis of call center arrival data: A bayesian approach. *Management Science*, 54(2):266–278.

Von Alan, R. H., March, S. T., Park, J., and Ram, S. (2004). Design science in information systems research. *MIS quarterly*, 28(1):75–105.

Zeltyn, S. and Mandelbaum, A. (2005). Call centers with impatient customers: many-server asymptotics of the m/m/n+ g queue. *Queueing Systems*, 51(3):361–402.

Zohar, E., Mandelbaum, A., and Shimkin, N. (2002). Adaptive behavior of impatient customers in tele-queues: Theory and empirical support. *Management Science*, 48(4):566–583.

Biography

Klas Gustavsson was born on the 14$^{\text{th}}$ of March 1988 in Östersund, Sweden. He received the Master of Science in Engineering from Mid Sweden University, Sweden in December 2015. He defended this thesis the 13$^{\text{th}}$ of June 2018 at Mid Sweden University in Sundsvall, to receive a Licentiate of Engineering degree.