Kingston University London

Kingston University London

School of Mechanical and Automotive Engineering

PhD THESIS

Academic Years 2011-2014

Ahmed A. Elfituri

Optimisation of Resources Deployment In A Call Centre By Using Stochastic Data In Simulation Models

Director of Studies:

Professor Andy W. Lung

October 2014

This thesis is submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

Acknowledgments

I would like to express my special appreciation and deep gratitude to my Supervisor Professor Andy Lung, who has been a tremendous mentor to me.

I would like to thank him for giving me a different perspective on things. The knowledge I gained during our work together is priceless. I thank him for the time and effort he so willingly contributed to this research, for allowing me to grow as a research scientist. His advice on both research as well as on my career has been priceless.

His desire to teach and to continue learning made this research both challenging and at the same time a fascinating experience, and for that I am truly grateful.

I would like to thank Jackie Deacon for her help and support.

A special thanks to my family. Words cannot express how grateful I am to my mother, father and my wife for all of the sacrifices that they've made on my behalf, supporting and encouraging me to attain my goals.

ABSTRACT

In recent years, call centres have been considered as an integral part of the modern businesses, since they play an important role in providing service delivery functions to their customers. A well-managed call centre, therefore, is crucial to ensure high level of customer satisfaction in today's competitive market. In order to achieve a high standard, managers of call centres face a very difficult set of challenges. At the top level, they must strike a balance between two powerful competing interests: low operating costs and high service quality. On a day-to-day basis, while simultaneously keeping low costs and high service quality, those managers must also employ appropriate techniques and tools in order to evaluate the true performance of their operations accurately. Such tools play a vital role in understanding the current system performance, evaluation of any proposed enhancement scenarios, and optimising operations management decisions under any unexpected operating conditions.

One of traditional operations management challenges for call centre managers is to tackle the multi-period human resources allocation problem. In this thesis, the staffing and staff scheduling decisions in single-skill inbound call centres were studied. These decisions are normally made under strict service level constrain in the presence of highly uncertain operations and demand of call centre services. Neglecting such uncertainty may lead to unrealistic decisions. The objective of this research thesis was to propose a framework to enhance the call centre performance through taking realistic optimal staffing and scheduling decisions. Realistic optimisation requires realistic modeling (evaluation) of call centre operations which is the main focus and contribution of this research.

The proposed framework has combined statistical, simulation, and Integer Programming (IP) techniques in achieving realistic optimisation. The framework begins by developing stochastic statistical data models for call centre operations parameters which are divided into service demand (arrival volumes) and service quality (service times, abandonment volumes, and patience time) parameters. These data models are then fed into a simulation model which was developed to determine the minimum staffing levels in daily an-hour periods. Finally, these staffing levels are considered as input to an IP model that optimally allocates the service agents to the different operating shifts of a typical working day. Application of the proposed framework to a call centre in Libya will also be presented to illustrate how its staffing and scheduling decisions could be improved by using the model.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

IVR	Interactive Voice Response
VRU	Voice Response Unit
CSR	Customer Service Representative
PSTN	Public Service Telephone Network
ANI	Automatic Number Identification
DNIS	Dialed Number Identification Service
PABX	Private Automatic Branch Exchange
ACD	Automatic Call Distributor
CTI	Computer-Telephone Integration
FIFO	First-In-First-Out
SL	Service Level
TSF	Telephone Service Factor
ASA	Average Speed Of Answer
QED	Quality Efficiency Driven
IP	Integer Programming
IID	Independent-Identically-Distributed
DES	Discrete Event Simulation

CHAPTER ONE

INTRODUCTION

CHAPTER ONE

INTRODUCTION

1.1 Background and Motivation

All modern businesses in today's competitive market include customers' satisfaction at the top of their list of priorities. Service businesses are interested in providing information and assistance, smoothly, to existing and prospective customers. During the last two decades, the reduced costs of telecommunications and information technology motivated the decision makers to consolidate information and service delivery functions to their running businesses, through call centres and their contemporary successors, contact centres. Those centres include a large number of resources allocated to provide prevalent means for modern businesses to communicate with their customers.

The economic importance of call centres is greater than many would think where the call centre industry is vast and rapidly expanding, in terms of both workforce and economic scope. In the United States of America, for example, customer service representatives occupy 2.3 million jobs – about 1.7 % of total USA employment – and this number is expected to grow by about 18 percent over the 2008-2018 period, faster than the average for all other occupations (United States Department of Labor, Acces date: November 13, 2012).

The trend in our economy from manufacturing towards services is well documented. One notable facet of this transition towards services has been the explosion of the call centre industry (Mehrotra & Fama, 2003). Call centres are an important function of most companies' day to day business activities.

Call centre research, while not as prevalent in earlier years, has grown in the past few years and has become a popular topic of discussion and research efforts in the industrial engineering and operations research fields. Common areas of call centre research include:

- Queuing theory
- Arrival models
- Workforce Management (WFM) models
- Simulation
- Integer Programming IP

1.2 Call Centre Definition

There is no universally accepted definition of "call centre" or "operator", although the following ones have been suggested: Call centre – a work environment in which the main business is conducted via the telephone whilst simultaneously using display screen equipment. This includes both parts of companies dedicated to this activity, such as internal helplines as well as whole companies. CC operator (also known as customer service advisor/ agent/handler) – is an individual whose job requires them to spend a significant proportion of their working time responding to calls on the telephone whilst simultaneously using display screen equipment. (www.hse.gov.uk/lau/lacs/94-1.htm), Acces date: May 24, 2012)..

A call centre is a business where the employees mainly handle incoming and/or outgoing telephone calls.
Typical services with outgoing calls are advertising campaigns, market research and selling by telephone. Examples of activities with incoming calls are customer services, giving information, taking orders and providing helpdesk functions. In the last few years operators have also started to handle e-mail, fax and SMS (short message service).

1.3 The history and development of the Call Centre business

Call centres have their origin in the USA, where they started in 1908 when it became possible to use the telephone to sell advertisements in a telephone book. In the beginning of the1960s Ford Motor Company started to search for possible buyers for their cars by making 20,000,000 phone calls to the consumers. For example, one of the largest telemarketing campaigns in Sweden was carried out in1978 when the Swedish telephone company (now Telia) decided to introduce the American concept "Yellow pages". Most of the advertisers had to buy their advertising space by telephone instead of being visited by a salesman. At the end of the 1980s the number of telemarketing companies started to grow, and more and more were established. The concept of call centre was born in the year 1991; in 1994 Telia introduced a campaign for call centres and the concept became firmly established in Sweden.
Call centres constitute one of the most rapidly growing businesses in the world. In the USA, call centres employ about 5 per cent of the workforce, while the figure is 1.3 per cent (2003) in Europe and 2 per cent in the UK (Data monitor, 1998, 1999).
Call centres are said to be the most rapidly growing form of employment in Europe today (Paul and Huws, 2002). Approximately 37 per cent of all new jobs within Europe during recent years have been in call centres.
It is estimated that, in South East Asia, India, the annual growth of call centres is 50 per cent.

1.4 What's new in a Call Centre?

Telephone operator work has become progressively more computerised over the last few decades, for example automatic distribution of calls and technical performance control has been introduced. These changes have resulted in a reduction of the variety of tasks performed by the operator, and increased repetitiveness and machine-regulation of the work. Computer-telephone interactive tasks, as performed in call centres are probably very special tasks.
Call centres use a range of information and communication technologies in order to maximise efficiency, and the technology that is the key to the call centre is the ACD (automatic call distributor). This computer directs the calls to the next available and logged-in operator. The computer also tracks how long it takes until the customer is connected, how long the call lasts

and the time that the operator not is working actively with calls or is disconnected because he or she has left the system. This eliminates the need for a central telephone operator by automatically processing the distribution of incoming telephone calls to the operators. Increasingly, ACD systems are connected to a range of databases using Computer Telephony Integration (CTI), which allows customer records to be transmitted to an agent's computer screen along with the call. In addition, many ACD systems have "voice response" mechanisms that are used to obtain basic information from the caller before they speak to an agent.

1.5 Importance of call Centre

The trend in our the economy from manufacturing towards services is well documented.
One notable facet of this transition towards services has been the explosion of the call centre industry (Mehrotra & Fama, 2003). Call centres are an important function of most companies' day to day business activities. They are often the link between a company and its customers and hugely impact the customer's perspective or point of view of a company. A call centre in the most general sense is a place, representing a business, which receives inbound calls from customers and/or makes outbound calls to customers, the latter usually being found in marketing and most commonly referred to as telemarketing.
Recently, over the past decade, the role of the call Centre has dramatically changed from simply handling calls into a complex, sophisticated environment, within many organisations and companies.

The primary function, of the majority of these centres, is to serve customers upon their requests through receiving telephone calls initiated by customers. Such centres are "inbound" call centres where they receive the service requests from the customers through the telephone calls only. In inbound call centres, the operations are highly labor-intensive where the staff members, who handle phone calls, are the main operational resource. The cost of those staff members typically comprises 60–80% of the overall operating budget of the call centres (Aksin, Z. et al.,2007).The quality of the service is typically viewed as a function of agents' performance that affects both how long the customer must wait to receive service and the value that the customer attributes to the information and service that is received. That is why the key factor in operations management and quality of service in call centres is the workforce management.
In a typical inbound call centre, calls arrive at random according to some complicated stochastic process(Avramidis,A.et al.2004), call durations are also random, waiting customers may abandon after a random patience time, some agents may fail to show up to work for any reason, and so on. All of these operational circumstances show an impelling need for effective management of the call centre resources; mainly its agents, under a high level of uncertainty. Furthermore, the call centre operations are always constrained by a minimum service level to be achieved. Service level (SL) is defined as the long-term fraction of customers whose waiting time for the service is no larger than a given threshold (e.g., a call centre may report the SL as the percent of callers who are answered within 20 seconds) (Gans, N. et al. 2003). Successful operations management requires that the decision makers work on achieving both low operating costs and high service quality simultaneously.

From the workforce management perspective, in order to meet these potentially conflicting objectives, it is extremely important to find a good match between the predicted workload and the scheduled workforce. In essence, inadequately sized workforce could lead to poor service quality (e.g. customers experience long waiting times). This could be avoided by scheduling a sufficiently large number of employees. However, it is undesirable to schedule too many employees because the economical objective of minimising operational costs should be also achieved. A small percentage of savings in workforce salaries simply means several million dollars (Bhulai, S. et al. 2008). This management perspective requires call centre managers to struggle with classical workforce management decisions including forecasting traffic, availability of agents, and deploying various resources (Aksin, Z. et al.,2007).

Upon the presented discussion regarding the size of the call centre industry and the complexity associated with its operations, call centres have emerged as a fertile ground for academic research. Moreover, this thesis is highly motivated by the promising future of this industry.

1.6 Functions of a Call Centre

There are six main operational functions in any type of call centre:

1- Workforce Management

Since an incoming call centre is at the mercy of incoming calls, the tasks associated with workforce management are among the most important functions of the centre. These tasks involve forecasting calls, calculating the optimal number of staff, creating work schedules, and managing daily service levels.

2- Quality Management

Due to the call centre's role as the primary point for customer communications, it is essential that these interactions be handled with utmost quality. The functions associated with quality management include customer surveying, call monitoring, performance assessment, and coaching.

3- Technology Management

The call centre today is filled with technology. From the moment the customer picks up the phone to the resolution of the call, many different technologies come into play. A critical function in every call centre is the effective management of these technologies including acquisition, implementation, and ongoing maintenance and management. These technologies are generally grouped as call delivery (including telecommunications infrastructure), call handling technologies, and call centre management tools.

4- Reporting and Communications

One of the components driving the call centre operation is information. An effective flow of communications is needed between customers and the centre, between the centre and other business units, and within the call center itself. Many different types of reports need to be generated every single day to show performance of the centre as well as that of individuals.

Therefore, reporting and communications is another essential function of call centre operations.

5- Financial Management

There is an array of costs associated with running a call centre operation. And some call centres may be generating revenue of their own as well. The aspects of creating and managing an operating budget and capital budget as well as using financial procedures to evaluate return on investment of proposed technologies or services are part of the financial management function of the call centre.

6- Risk Management

A final function within call centre operations is that of disaster recovery and contingency planning. Although not a task that is necessarily performed every day, it is a regular function within the best-run call centers to continually assess risks and potential solutions. (Reynolds, P.2003).

1.7 Problems associated with a typical call centre

Call centres are at the front line of customer service. But many things can stand in the way of optimum service provision.

According to some in-depth case studies, call centres are typically facing the following problems today;

1. Agent absenteeism

According to benchmarking firm Dimension Data, the average annual absence rate in call centres across the globe is 11%. To the layperson, this might not seem particularly high. But the stark reality is that a 100-seat contact centre with 11% absenteeism will only have an average of 89 seats occupied at any one time.

Unsurprisingly, a shortfall of this magnitude can have a huge impact on quality of service. Because there are fewer staff available to handle customer interactions, wait queues tend to increase and agents are put under pressure to spend less time on each call. Over extended periods of time, absenteeism can impact on staff morale and may even foster similar behaviour in those left to 'carry the can'.

2. Customer churn

Customer attrition is a huge problem, with research from (Genesys-EMG, Alcatel-Lucent) revealing that a massive 73 per cent of UK consumers have – at one point or another – eliberately chosen to end their relationship with a goods or service provider.

The cost of such losses can be enormous. Indeed, Genesys estimates that UK businesses lose approximately £15.3billion every year because customers have either chosen to abandon the purchase they were originally making or have defected to a competitor.

3. Forecasting accuracy

Forecasting accuracy – better described as forecasted contact load vs. actual contact load is a performance metric that reflects the percent variance between the number of inbound customer contacts forecasted for a particular time period and the number of said contacts actually received by the center during that time. It is a critical, high-level objective in all Call Centre environments.

Underestimating demand leads to understaffing. This, in turn, leads to long wait times in queues, frustrated customers, burned-out agents and high toll-free costs (due not only to the long hold times, but also to the longer call times that might result from dedicating a portion of the call to caller complaints about hold times). However, overestimating demand results in waste, overstaffing and increased idle time.

Forecasting accuracy should not be reported as a summary of forecasted versus actual contacts across a day, week or month, but rather as an illustration of accuracy for each reporting interval, typically half-hours.

4- Call Abandons

Call centers measure the number of abandons as well as the percentage of calls that abandon, called the abandon rate or abandonment rate. This abandon rate can translate into lost customers, so it is important to track it to identify patterns of abandon behaviors.

There are also a number of generic measures, these include:

- Total calls answered – the overall volume of calls the centre was able to handle
- Total calls abandoned – the number of customers who hang up before being answered as a % of total calls made.
- Service level – the % of calls answered within a specified time frame – usually something like 90% within 20 seconds
- Average call duration – the overall talk and wrap time for the calls. (Alex, C. 2010)

All of the above measures are important in how well the call centre will perform overall. If there is a high volume of calls but not enough resource (for whatever reason) to answer them, then more customers will abandon and the service levels will not be met. For the call centre Manager it can be a real juggling act to meet both the demands of the customer, the financial restraints they operate under and ensuring that the employees are treated according to their terms and conditions.

1.8 Research Question and Thesis Contribution

The aim of this work is to improve the staffing and scheduling decisions by developing a stochastic data model, simulation model to determine the minimum number of staffing levels, optimum combination of shift scheduling and achieving the minimum cost of the call centre.

The objective of the current work is to propose a framework to enhance the call centre performance through taking realistic optimal staffing and scheduling decisions. Realistic optimisation requires realistic modeling (evaluation) of call centre operations which is the main contribution of the current work.

In order to achieve that, the special features characterizing the call centre system should be considered. General characteristics include that call centre systems are complex, and most of their operational parameters exhibit highly stochastic nature and require some form of optimisation such as in workforce management decisions. Traditional analytic queuing models and optimisation methods, in contrast, only apply if the systems are sufficiently simple and simplifying assumptions are made. Thus, a combination of simulation and optimisation is more beneficial in developing the proposed performance enhancement framework.

The first step in the proposed framework is to develop appropriate stochastic models for the various operational parameters; namely, arrival counts, service times and the two closely related parameters of abandonment rates and patience times. The next step is to build a simulation model which captures the main characteristics of the call centre dynamics based on the previously developed stochastic models for the various input parameters. The simulation model is in turn used to determine the minimal staffing levels in each daily one-hour period. Finally, after an appropriate staffing level has been determined for each period, this time-varying staffing is fed into an integer linear programming model to choose, from all feasible shifts covering the staffing requirements, the solution achieving the minimal cost.

The developed model will be tested in a call centre in Libya, and the results were very encouraging in terms of improving the performance prediction will be analysed to increase the service level and reducing the abandonment calls.

1.9 Scope

It is very common in call centre literature to distinguish four phases in the process of workforce management as described in (Bhulai, S. et al. 2008) and (L'Ecuyer, P. 2006):

Phase I. *Workload prediction:* the process that is concerned with the prediction of the future amount of service demand arriving to the call centre.

Phase II. *Staffing:* the problem in which the predicted workload capacity is translated into numbers of required agents such that a predetermined service level is met.

Phase III. *Shift scheduling:* the determination of how many agents should be assigned for each possible work schedule for the day or week.

Phase IV. *Rostering:* the pairing of shifts into rosters and the assignment of employees to the rosters.

The current work is dealing with phases II and III (multi-period human resources allocation problem) of the process of workforce management in call centres. It tackles the staffing and scheduling problems due to their strong impact on both operating costs and service quality. Moreover, it aims to deal with some of several shortcomings in the existing solutions of both problems in literature.

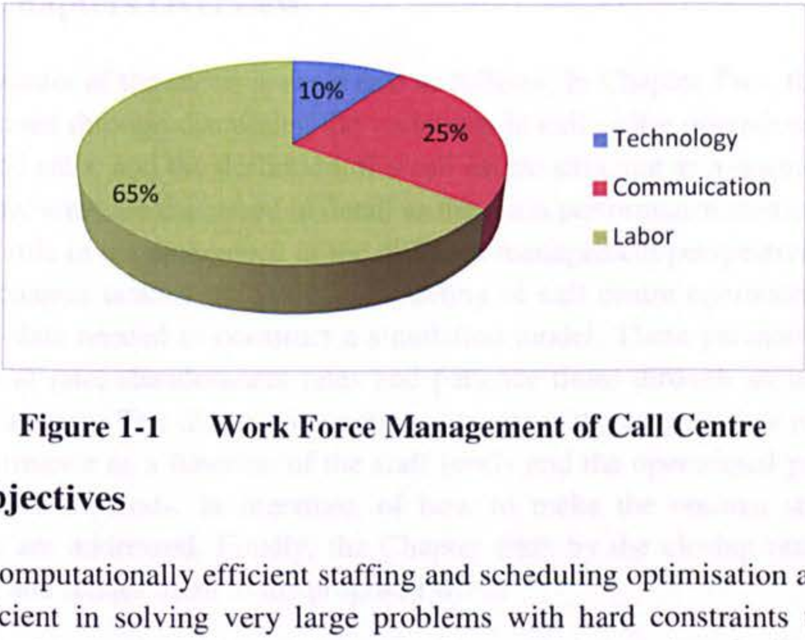


Figure 1-1 Work Force Management of Call Centre

1.10 Objectives

Having computationally efficient staffing and scheduling optimisation algorithms will make it very efficient in solving very large problems with hard constraints to optimality or near-optimality in short period's time. However, the computational efficiency is not the sole objective in solving the staffing and scheduling problems in call centres.

A more important objective to be considered is how realistic the obtained optimal solutions are. The extent to which the optimal solutions obtained by different optimisation algorithms are realistic depends mainly on the quality of inputs used in running these algorithms. In the staffing and scheduling problems context, the staff levels and schedules depend on two sets of parameters, which are the service demand (arrival volumes) and service quality (service times, abandonment volumes, and patience time) parameters, as mentioned above. These parameters exhibit considerable level of uncertainty. Traditionally, the optimisation models in literature use just simple averages to represent these parameters. This may result in a probably optimal solution which could be unrealistic at all. Therefore, the objective of the current work is to propose a framework to enhance the call centre performance through obtaining realistic optimal staffing and scheduling decisions.

This framework is developed by combining statistical modeling, simulation, and IP techniques in three consecutive steps:

I. ***Developing stochastic data models*** for call centre dynamics parameters. This requires extensive study and analysis to develop valid realistic data models to avoid the shortcomings due to neglecting the call centres highly stochastic nature.

II. ***Developing a valid discrete-event-simulation model*** of the call centre dynamics that mimics its stochastic operations through using the developed stochastic data models. This valid model is used in evaluating the current service levels to determine the enhancement goals. Afterwards, the simulation model is run several times to determine the optimal staff levels in each daily period in an iterative manner.

III. ***Developing an integer programming model*** which optimally allocates the service agents, determined in step II, to the different operating shifts of the working day.

1.11 Chapters Overview

The remainder of the thesis is organised as follows. In Chapter Two, the call centre dynamics are examined through discussing the variations in call centre operations, the handling method of inbound calls, and the definition of a call centre structure as a queuing system. The service quality measures are discussed in detail as the main performance measures used in call centres and their role in the emergence of the different management perspectives of call centres. After that the chapter tackles the realistic modeling of call centre operational parameters, namely the input data needed to construct a simulation model. These parameters include the service rate, arrival rate, abandonment rates and patience times through an in-depth analysis of the related literature. The discussion continues to cover the area of how to analyze and evaluate the performance as a function of the staff levels and the operational parameters. Afterwards, the different methods, in literature, of how to make the optimal staffing and scheduling decisions are addressed. Finally, the Chapter ends by the closing remarks on the reviewed literature and relates them to the proposed work.

Chapter Three presents the proposed methodology to address the three stages involved in solving the workforce management in a call centre. The stochastic data models development stage is covered; where two important input parameters are examined: the arrival counts and the customer's abandonment.

In Chapter Four, the methodology to conduct the simulation model is presented and it is run using the developed data models. The chapter focuses on a proposed methodology to conduct reliable validation for the developed simulation model. The optimisation models are discussed how they are used for tackling both staffing and scheduling problems.

Chapter Five demonstrates the application of the proposed stochastic data models on real call centre data. This chapter presents a detailed prelude data analysis step for each studied parameter determining the characteristics of that parameter which led to choosing a specific data model for its representation. The techniques for estimating the parameters involved in each data model and the model validation are discussed. The simulation modeling and its related analysis of the real call centre are studied. The details of the simulation model construction phase is discussed according to the proposed methodology presented in Chapter Four and the results of running the simulation model are examined. The optimisation step is applied to solve the staffing and scheduling problems and the results are presented.

Chapter Six includes the discussion and the implications of the research.

Chapter Seven summarizes the main conclusions of the proposed work and identifies future research directions.

CHAPTER TWO

LITERATURE REVIEW

CHAPTER TWO
LITERATURE REVIEW

2.1 Introduction

2.1.1 Overview of Call Centre Operations

Call centres represent a critical component of the worldwide services infrastructure. A call centre is operated through several resources including personnel, computers and telecommunication equipment to deliver telephone-based services. The working environment of a large call centre could be envisioned as an endless room in which people with earphones sit in front of computer terminals, providing tele-services to customers acquiring these services.

Call centres could be categorized along many different dimensions. They vary according to the type of provided services from customer service, help desk, and emergency response services, to tele-marketing and order taking. They also vary in terms of size and location from small local sites such as emergency response services, (Matteson, D. et al. 2011), to large international centres such as international cellular phone companies. Furthermore, the latest telecommunications and information technology allow a call centre to be the virtual embodiment of a few or many geographically dispersed operations (Kalaignanam, K. and Varadarajan, R., 2012).

The organisation of work may also vary dramatically across call centres. Two different types of call centres have emerged based on the required human resources skills. The first type is the single-skill call centre where the skill-level required in handling calls is low, a centre may train every employee to handle all call types, and calls may be handled on a first come, first-served basis (Roubos, A. et al. 2011). On the other hand, multi-skill call centres appear when more highly-skilled operators are required (Bhulai, S. et al. 2008). Each operator may be trained to handle only a subset of the different call types, and *skills-based routing* (SBR) may be used to route calls to appropriate agents. In turn, the organisational structure may vary from the very flat to multi-layered structures. In flat structures, all agents are essentially exposed to all incoming calls. In the multi-layered structures, each layer represents a level of expertise where customers may be transferred through several layers before being completely served.

Another important classification criterion is the call initiation origin; whether it is from the customer or from the call centre. Call centres in which calls are initiated by outside customers seeking the centre services are called "inbound" call centres (Roubos, A. et al. 2011). Inbound call centres represent the majority of existing call centres; these types of centres provide customer support, help desk services, reservation and sales support for airlines and hotels, and order-taking functions for catalog and web-based merchants. When calls are initiated from the call centre then the centre is called an "outbound" call centre where it handles outgoing calls to selected pools of customers such as in tele-marketing and survey businesses (Pichitlamken,

J. et al. 2003). Recently, some inbound centres incorporate outbound functions to its services to high-value customers who have abandoned the system before being served (Gans, N. et al. 2003).

Many inbound call centres use *interactive voice response* (IVR) units, also called *voice response units* (VRUs) in addition to *customer service representatives* (CSRs) in providing services to their customers (Tezcan, T. and Behzad, B. 2012). IVR units are specialized computers that allow customers to communicate their needs and to "self-serve." Customers interacting with an IVR use their telephone key pads to provide information, such as account numbers or indications of the type of service desired. In response, the IVR uses a synthesized voice to report information, such as bank balances or departure times of planes.

Contact centres are the contemporary successors of the call centres. A contact centre is a service centre in which services could be presented through other communication media in addition to agents and IVRs, such as email, fax, web pages, or chat. The trend toward contact centres has been stimulated by societal hype surrounding the internet and by customer demand for channel variety, as well as by the potential for efficiency gains. In particular, requests for email and fax services could be "stored" for later response and it is possible that, when standardized and well managed, they could be made significantly less costly than telephone services (Gans, N. et al. 2003).

.

2.1.2 Inbound Calls Handling

The large-scale emergence of call centres has been enabled by technological advances in information and communications systems. These technologies and how they function are illustrated in Figure 2-1. The process by which a call centre serves an incoming call could be described as a sequence of operations which starts by connecting the customer to the centre and ends with aborting this connection by the customer.

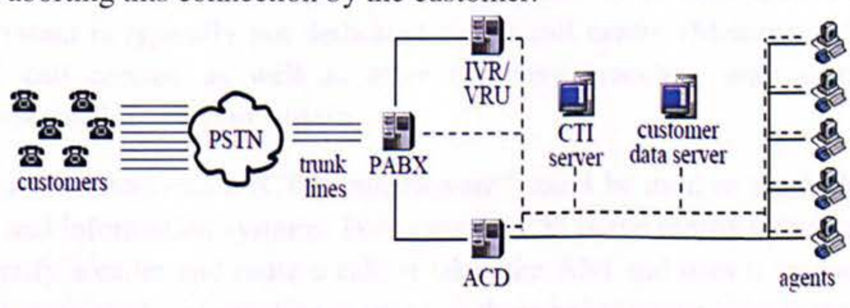


Figure 2-1: Schematic Diagram of Call Centre Technology (Gans, N. et al. 2003)

The customers begin the process of acquiring the call centre service by calling the call centre phone number. The long-distance or *public service telephone network* (PSTN) company, that provides the telephone service to the call centre, needs to define two vital pieces of information about each call; the number from which the call originates, often called the *automatic number identification* (ANI); and the number being dialed, named the *call's dialed number identification service* (DNIS).

Each call centre has its own, privately-owned switch, called a *private automatic branch exchange* (PABX), and the caller's DNIS locates the PABX on the PSTN's network. If the call centre has more than one location on the network – all reachable via the same telephone number – then a combination of the ANI (which gives the caller's location) and the DNIS may be used to route the call. Conversely, more than one DNIS may be routed to the same PABX.

Each call centre has a very important capacity resource other than the CSRs. This resource is a limited number of telephone lines, often called trunk lines, used to connect the PABX to the PSTN. If there are one or more trunk lines free, then the call will be connected to the PABX. Otherwise, the caller will receive the well-known busy tone. If the call is connected it may be served in a number of phases.

- Calls may be connected through the PABX to an IVR that queries customers on their needs.
- Customers may need or desire to speak with a CSR to get more details than obtained using the recorded messages, and in this case calls are handed from the IVR to an *automatic call distributor* (ACD). An ACD is a specialized switch designed to route calls, connected via the PABX, to any suitable idle CSR within the call centre.

The idleness of one or more trunk lines does not necessarily mean that there are idle agents to handle the incoming customer. In this case, customers are put on hold and typically exposed to music, commercials, or other information. Customers on hold may become impatient and hang up before they are served. When they hang up, they are said to abandon the queue or to renege. Customers who are patient enough and willing to wait will be connected to a CSR after a variable waiting time. Once a customer starts to speak with the agent, the customer communicates his/her need to the agent. At the same time, agents work via a PC or terminal with a corporate information system to process the required service or to obtain the required information by the customer. In large companies, such as airlines and retail banks, the information system is typically not dedicated to the call centre (Mehrotra, V. et al. 2012). Rather, many call centres, as well as other company branches, may share access to a centralized corporate information system.

Computer-telephone integration (CTI) "middleware" could be used to more closely integrate the telephone and information systems. For instance, CTI is the means by which a call's ANI is used to identify a caller and route a call: it takes the ANI and uses it to query a customer database in the company's information systems; if there exists a customer in the database with the same ANI, then routing information from that customer's record is returned. The routing information would be the customer's preferred language.

Finally, the service needs not end with the call. Callers who are blocked – due to the unavailability of any idle trunk lines – or abandon – due to waiting for time longer than his/her patience – may try to call again, in which case they become retrials. Callers who speak with CSRs but are unable to resolve their problems may also call again, in which case they become returns (Artalejo, J. and Pla, V. 2009).

2.1.3 Call Centres as Queuing Systems

The structure of a call centre service system and the way in which it interacts with customers could be represented as a typical queuing system.

Figure 2-2 shows a general operational scheme of a simple call centre as queuing model. According to this general call centre structure, the call life span within a call centre consists of several consecutive stages and it may end at any stage. First, the call arrives in the call centre that has a set of K trunk lines that is determined according to a contract with the PSTN company. If there is no trunk line available, the call is blocked from entering the queuing system and the caller receives the well-known busy signal. Otherwise, the call seizes one of the available trunk lines. At this instant, the ACD checks whether any of the available N agents ($N \leq K$) are idle or not. If there is any idle agent, the ACD connects the caller directly to that agent and the service process starts directly. Otherwise, the caller will begin to wait in a tele-queue for an agent to become available.

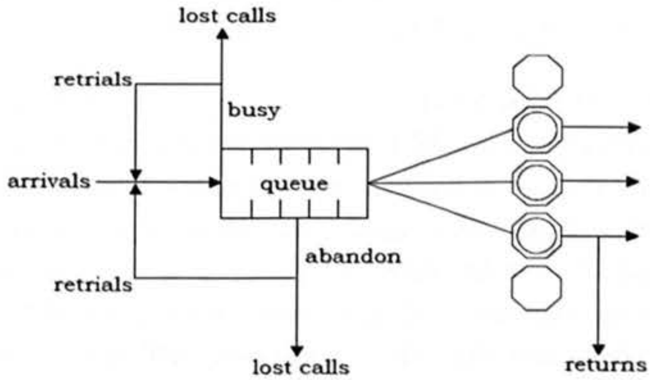


Figure 2-2: A general operational scheme of a simple call centre as queuing model (Gans, N. et al. 2003)

While waiting, the caller may become impatient and abandons the queue before being served, especially, if he/she experiences relatively long waiting time. If the caller is patient and remaine in the queue, he/she would eventually reach the front of the queue and then obtain the acquired service. In simple call centres, the service discipline is usually *first-in, first-out* (FIFO). The callers who do not receive the service – blocked or abandoned – may become retrials that attempt to acquire the service again. Those who do not retry are considered lost calls. On the other hand, callers who obtained their acquired service may become returns either to acquire a new service or to investigate about problems regarding the already obtained service.

2.1.4 Operations Performance Measures

The performance of the call centre service system is evaluated through a set of output parameters, as documented in (Robbins, T. and Medeiros, D. 2006), such as the service level, the average speed of answer, the abandonment percentage, and agent occupancy. The most frequently used measure in practice is ***Service Level*** **(SL)** that is also called the *telephone service factor* (TSF), SL is the fraction of calls for which the delay is below a specified level. For example, a call centre may report the SL as the percent of callers who are answered within 20 seconds. The most common service level is 80/20 which refers to 80% of incoming calls answered within 20 second. The SL could be measured and controlled separately per time period (30 minutes, hour, etc.) and by call type, or in an aggregated way either over a period of time (daily, weekly, etc.) or for all call types. An important motivation for studying this measure is that for many types of call centres that provide services, in several countries, there are governmental regulations on the minimal acceptable SL and the call centres may have to pay very large fines when this SL is not met (L'Ecuyer, P. 2006).

Average Speed of Answer **(ASA)** is another performance measure that is the average time callers spend on hold, waiting to speak to an agent (CSR). It is the average waiting time in the agents' queue calculated for only customers who succeed in acquiring service. Due to the fact that ASA does not include the time that abandoned calls spend waiting, a reasonably full picture of congestion requires, at a minimum, both ASA and abandonment percentage statistics (L'Ecuyer, P. 2006). Callers that are put on hold and hang up while in queue are said to have abandoned the system. The proportion of all calls that abandoned is known as the ***Abandonment Percentage*** **(Aband %)** and it is a key metric in most call centres. This metric could also be per call type, per period, and aggregated.

Another important performance metric is the ***Agent Occupancy*** which is a measure of agent utilization that excludes the time that agent is unavailable to serve calls. It also may be measured per agent type and per period. In theory, they would like to have their agents occupied as much as possible under the quality of service constraints. But important human factors must also be taken into account. Overstressed agents tend to perform more poorly in terms of both quality and speed. Generally, it is a bad idea to have occupation ratios above 90–92% for a sustained period of time.

2.1.5 Operations Management Perspectives

The essence of call centre operations management is the relationship between queuing behavior and staff level decisions. Thus, the fundamental management tradeoff is between service quality and operational costs. Performance analysis supports this tradeoff by calculating the attained service level and the resource occupancy/utilization as functions of offered load and available resources (Koole, G. and Mandelbaum, A. 2002). The common management regimes in call centres are classified into three basic categories that describe the staffing/customer service objectives of the call centre.

Those regimes are described as follows (Robbins, T. and Medeiros, D. 2006):

1. ***Quality Driven Regime.*** The regime in which customer waiting costs are assumed to dominate the cost of capacity and the objective is to serve the majority of customers without delay. Staffing levels are increased linearly with the offered load. The average utilization in this regime is typically low, on the order of 65-75%, and the average customer waiting time is also low.

2. ***Efficiency Driven Regime.*** Staffing costs are assumed to dominate the cost of customer delay and the operational objective in this regime is to maximize the utilization of the staff.

3. ***Quality Efficiency Driven* (QED).** An operational environment that attempts to strike some balance between efficiency and customer service. Unlike the quality regime where the fraction of delayed customer is near zero, or the efficiency regime where the fraction delayed is near one, the QED regime balances costs and attempts to achieve some steady delay proportion between 0 and 1.

2.2 Stochastic Input Data Analysis and Modeling

The current work is concerned with the determination of the optimal number of agents serving the customers during the different daily periods under strict performance measures. From the queuing theory perspective, performance measures and the corresponding server capacity are determined according to a set of input parameters. Input parameters include the demand on the offered services from the call centre (the arrival counts), the handling duration of each acquired service (service times), callers' patience times, and abandonment rates.

The input parameters show a lucid nature where they experience significant day-to-day variations, and seasonality patterns over a time scale of weeks and months. Ignoring this fact, concerning the input parameters, results in significant inaccuracy. Such inaccuracy would result in wrong performance estimates and thus wrong staffing decisions. Staffing decisions (the problem in which the predicted workload capacity is translated into numbers of required agents such that a predetermined service level is met) under such wrong estimates, would, for example, result in flat staff levels over the time. Pursuant to this, the call centre performance under those inaccurate estimates would be affected negatively. This is due to the fact that any variation in those input parameters especially during peak load periods would lead to the customers encountering significant delays and eventually abandoning without acquiring any service (Robbins, T. and Harrison, T. 2008).

Accurate input data models development due to the stochastic nature of input parameters is an essential requirement to make accurate staffing decisions. In essence, accurate staffing decisions should avoid either understaffing that leads to unsatisfactory performance or overstaffing that leads to high agents' costs. Accurate estimation of those parameters is a corner stone in building robust models, to be used as a reliable decision making tool, taking into consideration that the call centre performance is immensely sensitive to any variation in those parameters especially in peak load periods (Gans, N. et al. 2003).

2.2.1 Modeling Approaches

The analysis and modeling of call centre input data could be performed using three different approaches (Koole, G. and Mandelbaum, A. 2002):

I. ***Descriptive models*** provide summaries of the empirical data obtained from the real system in the form of tables and histograms such as a histogram of total daily number of arrived calls as the call arrivals fluid-like model presented in (Mandelbaum, A. et al. 2002).

II. ***Explanatory models*** utilize regression and time series analysis to determine and describe the desired parameters using explanatory variables; for example, a type of first-order autoregressive structure is suggested by (Brown, L. et al. 2005) for the random daily effects influencing the daily call volumes.

III. ***Theoretical models*** provide a mathematical representation to fit the empirical data using theoretical statistical distributions and to sample from them. The arrival counts to a call centre, for example, are modeled using a Poisson mixture model by (Jongbloed, G. and Koole, G .2001).

2.2.2 Modeling Issues

In spite of the high ability of modern computerized data collection systems used in call centres, the developing of accurate data models faces, in practice, a wide range of real world problems which complicate this task considerably. The level of detail, at which the operational data is collected, is limited to aggregate data parameters over fixed short time intervals varying between fifteen to sixty minutes, hence, the call-by-call data is not available. The available data, for example, is the number of arrived calls to the centre in each time interval, and the average service time in that interval. Moreover, the data regarding the customer patience time is very poor; the number of callers who abandon the queue after fixed time epochs, where common choices are 10, 20, 30, and 60 seconds, is the only in-hand detail about customer patience, while a statistical model representing the time that a customer will wait before deciding to abandon the queue needs to be developed. The main issue with such kind of data is that standard parameter estimation methods cannot be used, to estimate the required statistical models, due to the unavailability of records for each individual call (Henderson, S. 2003). Consequently, developing a suitable data modeling method that could deal with this issue is not an easy task.

In addition to the lack of call-by-call data issue, the stochastic nature of call centre operations complicates the data analysis and modeling because of the high level of uncertainty in the real data. Call arrivals, for example, vary significantly from time period to another time period of the same day above and beyond the weekly and monthly seasonality. This amount of uncertainty makes the estimation problem more complicated and requires extensive and careful data analysis. The development of call centre data models is also affected by several human factor issues that arise due to the behavior of call centre agents. It is very likely in call centres to find that the number of ready agents, to answer any incoming call immediately at any instance, is less than the number of scheduled agents (*Logged-on agents*). This shortage may be due to several reasons such as sudden absenteeism, unplanned coffee breaks, and trips to lavatories, unplanned breaks because of health issues such as sleeping disorders, and eyesight problems, inter-personal relationships between agents leading to friendship breaks, and so on. The issue here is how to quantify the amount of such shortage occurring due to human aspects.

In this work, both descriptive and theoretical modeling approaches are used to study and develop suitable stochastic data models for the arrival counts to inbound call centres. In the model development, all issues mentioned above are considered to increase the validity of the developed models as the model accuracy is a function of its building blocks validity.

2.3 Input Data Modeling in Literature

In the work of (Gans, N. et al. 2010), a call centres could be treated as an Erlang-C (M/M/s) multi-server model. That model is based on three main assumptions. First, the incoming customers are considered homogenous (i.e., the call arrivals follow a Poisson process). Second, the servers are also considered to be homogeneous which means that service-times are independent and identically distributed according to an exponential distribution. Third, the customers have infinite patience; hence, they never abandon the queue (Robbins, T. 2010). Modeling call centres according to these assumptions is useless as they are not capable to touch upon any of the characteristics of the real data. In literature, several statistical models have been considered to deal with the modeling issues discussed above. Those models cover the different input parameters.

2.3.1 Arrival Counts Models

There are several problems encountered with call centres data especially with arrivals data. The stochastic nature of arrival counts exhibits three main properties that characterize the arrival process in telephone inbound call centres (Avramidis, A. et al. 2004):

Property A. The call volumes vary significantly over the time (*arrival rate uncertainty*).

Property B. The variance of the arrival volumes is much greater than their means (*data over- dispersion*).

Property C. The dependence between arrival volumes in the successive periods of a day is very strong (*strong positive correlation*).

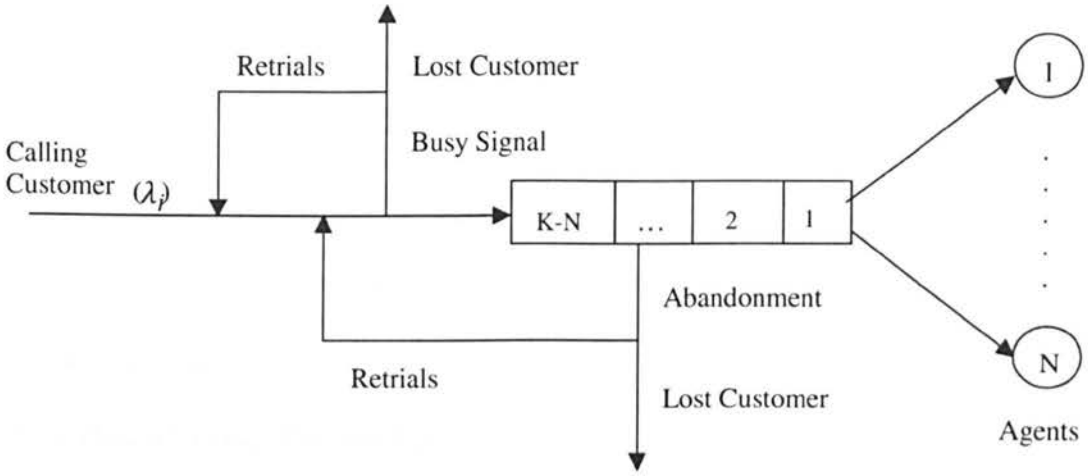


Figure 2-3: Queuing System of the Call Centre

The Poisson Process

According to the Palm-Khintchine theorem, the counting process of events occurring from a large number of independent sources – where anyone of the sources contribution to the total number of events is small – behaves asymptotically as a Poisson process. In call centres context, when a large number of independent customers, each of whom has a comparatively small calling probability, are possible callers of a call centre, the Palm-Khintchine theorem provides justification for using the Poisson process to model the arrival process to that call centre. The Poisson process is the most elementary random process used in modeling the arrivals of customers to call centres. The standard definition of this process as it appeared in (Ross, S. 2007) is as follows.

Definition 1. The counting process $\{N(t), t \geq 0\}$ – where N(t) is the total number of "events" that occur by time t – is said to be a Poisson process having rate λ, $\lambda > 0$, based on the following assumptions:

i. $N(0) = 0$.

ii. The process has independent increments. That is, for all s, t, v, and $u \geq 0$, $N(t) - N(s)$ and $N(u) - N(v)$ are independent for any non-overlapping intervals $(s, t]$ and $(v, u]$.

iii. The process has stationary increments. The number of events in any interval of length s is Poisson distributed with mean λs. That is, for all $s, t \geq 0$

$$P\{N(t+s) - N(t) = n\} = e^{-\lambda s}\frac{(\lambda s)^n}{n!}, \qquad where\ n = 0, 1, \ldots \qquad \text{2-1}$$

It must be noted that property A of the arrival process in a call centre contradicts the assumption of the standard Poisson model that the arrival process has stationary increments with the same arrival rate as the arrival rate varies considerably during different daily periods. Moreover, the arrival rate – in the presence of property $\boldsymbol{B}$ – could not be modeled using a Poisson distribution as it assumes that the mean and variance are equal while the data experiences considerable over-dispersion. Additionally, the independent increments assumption is inconsistent with property $\boldsymbol{C}$ where the strong positive correlation refutes the independence between the non-overlapping periods. Consequently, these properties render the standard Poisson process model inadequate.

The Non-Homogeneous Poisson Process

In view of property A, the problem of uncertain arrivals during different periods of the day could be solved by considering the arrival process during separate periods, for instance one-hour periods, as a Poisson process but with a rate varying from a time-period to another. This time sampling Poisson process generates a non-homogeneous process that, by definition, allows to model time-dependent arrivals (Ross, S. 2007).

Definition 2. The counting process $\{N(t),\ t \geq 0\}$ – where $N(t)$ is the total number of "events" that occur by time t – is said to be a *non-homogeneous Poisson process* (*NHPP*) with intensity function for time varying arrival rate $\lambda(t)$, $\lambda(t) > 0$, based on the following assumptions:

i. $N(0) = 0$.

ii. $\{N(t),\ t \geq 0\}$ has independent increments. That is, for all s, t, v, and $u \geq 0$, $N(t) - N(s)$ and $N(u) - N(v)$, $(s, t]$ and $(v, u]$.

iii. The process has stationary increments. The number of events in any interval of length s is Poisson distributed with mean $s\lambda(t)$. That is, for all s, $t \geq 0$

$$P\{N(t+s) - N(t) = n\} = e^{-s\lambda(t)} \frac{(s\lambda(t))^n}{n!}, \qquad \text{2-2}$$

$$where\ n = 0, 1, 2, \ldots$$

The main problem here is how to estimate the time-varying arrival rate. The arrival rate function is assumed, in (Henderson, S.2003), to be piecewise constant over the subsequent time-intervals of the day and that could be estimated from the data; this estimator could be a consistent estimator of the original arrival-rate function by performing an asymptotic analysis of this method. In (Massey,W. et al.1996), a piecewise linear rate function is proposed for the arrival rate and then several ways are investigated for the parameters estimation of this model such as ordinary least squares (OLS), iterative weighted least squares (IWLS) and maximum likelihood (ML) methods. A piecewise polynomial approximation is suggested in (Kao, E. and Chang, S.1988) to represent the rate function using maximum likelihood estimators. Time-inhomogeneous Poisson processes successfully address the problem of daily period-to-period variability; the other levels of variability (day-to-day variability as well as weekly and monthly variability) could be handled by providing a separate model for each period in which the arrival patterns are consistent by means of clustering various periods.

The Doubly Stochastic Poisson Process

According to the above discussion, ignoring any of the calls arrival process main properties will produce inaccurate queuing/simulation models that are to be used later in solving the staffing problem which may bring the validity of this work into question. If the variability in the real arrival process is higher than that obtained by the standard Poisson process, then the estimated service level is lower than it would be otherwise (Steckley, S.2005). In order to consider the aforementioned data properties in the data generation model, the conditional Poisson process has been suggested for use in modeling the call arrival counts.

Definition 3. Let $\{X(t), t \geq 0\}$ be a counting process and there is a positive random variable Λ such that conditional on $\Lambda = \lambda$ the counting process is a Poisson process with rate λ. That counting process is called a conditional Poisson process representing the doubly stochastic arrival model. This model stands for the call centre context when the following hold:

i. If the random vector of arrival counts is $\mathbf{X} = (X_1, X_2 \ldots, X_k)$, where X_i is the number of call arrivals in period i and k is the number of daily periods, then the random arrival counts X_i follow Poisson distributions with probability mass function

$$P(X_i = x) = e^{-\Lambda_i} \frac{\Lambda_i^x}{x!} \qquad \text{2-3}$$

ii. The rate Λ_i of a Poisson random variable X_i is a random variable generated randomly from a period-dependent distribution for Λ on $(0, \infty)$. This accounts for the over-dispersion problem (property ***B***) in the standard Poisson model (Jongbloed,G.and Koole,G.2001).

iii. The Poisson random variable X_i with a rate Λ_i is generated separately for each daily period from a separate standard Poisson process and this accounts for the high level of variability in arrival counts data (property ***A***).

iv. There are many ways to estimate the rate of the doubly stochastic Poisson process such as those proposed in (Jongbloed,G. and Koole,G.2001) and (Avramidis, A.et al.2004). In the present study a new model for the rate function – daily proportion-based arrival rate – is developed, studied and compared to the gamma dependent arrival rate, discussed below. The proposed model will be discussed in chapter 3.

According to (Avramidis et al.2004), the arrival rates Λ_i are modeled as dependent random variables randomized by a common gamma variable that accounts for the correlation between the number of arrivals in the subsequent periods (property ***C***) where

$$\Lambda_i = W\lambda_i, \qquad W \sim Gamma(\gamma, 1)$$

Their proposed function of arrival rates results in a *negative multinomial distribution* for the vector $\mathbf{X}$ with parameters $(\gamma, \lambda_1, \lambda_2 \ldots \lambda_k)$ where the probability mass function is given by

$$f(\mathbf{x}) = \frac{\Gamma(\gamma + \sum_{i=1}^{k} x_i)}{\Gamma(\gamma)\prod_{i=1}^{k} x_i!}\left(\frac{1}{1+\sum_{j=1}^{k}\lambda_j}\right)^{\gamma}\prod_{i=1}^{k}\left(\frac{\lambda_i}{1+\sum_{j=1}^{k}\lambda_j}\right)^{x_i} \qquad 2\text{-}4$$

The parameters of the *negative multinomial distribution* could be estimated by the maximum likelihood estimation method (MLE). If the vector $\left[\mathbf{X}_j = \left(X_{1,j}, X_{2,j}, \ldots, X_{k,j}\right)\right]_{j=1}^{n}$ represents the data set of arrival counts observations – where n is the number of similar days in the data set and k is the number of daily periods – then the maximum likelihood estimators (MLEs) could be obtained according to the following estimation algorithm:

$$Daily\ Call\ Volume\ Y_j = \sum_{i=1}^{k} X_{i,j}, \qquad for\ j = 1, 2, \ldots, n \qquad 2\text{-}5$$

$$F_l = \frac{1}{n}\sum_{j=1}^{n} I\{Y_j \geq l\}, for\ l = 1, 2, \ldots, M \qquad where\ M = max(Y_j) \qquad 2\text{-}6$$

$$\sum_{l=1}^{M} (\hat{\gamma} + l - 1)^{-1} F_l = log\left(1 + \frac{1}{n\hat{\gamma}}\sum_{j=1}^{n} Y_j\right) \qquad 2\text{-}7$$

$$\hat{\lambda}_i = \frac{\sum_{j=1}^{n} X_{i,j}}{n\hat{\gamma}}, \qquad for\ i = 1, 2, \ldots, k \qquad 2\text{-}8$$

Y_j and F_l are calculated directly from the observed data. $\hat{\gamma}$-value is calculated by solving equation 2-7 above numerically, and then $\hat{\lambda}_i$ is obtained for each period. This estimation process is repeated, separately, for each day within the week due to the high amount of day-to-day variability (e.g., $\hat{\lambda}_1$ on Saturdays is different from $\hat{\lambda}_1$ on Sundays and so on).

2.3.2 Service Time Models

In large call centres (with thousands of agents), the economic saving of decreasing (or cost of increasing) the average service time by even one second could be millions of pounds annually (Gans, N. et al.2010). Thus, the service times are worthy of high attention from the practical view-point. This practical consideration also has a direct impact on the theoretical methodology adopted. In essence, any misrepresentation in the statistical models of service times may lead to strong doubt in its validity.

According to the Erlang-C model, the service time is assumed to follow an exponential distribution. The main reason is that the calculations required to characterize the system's performance is greatly simplified due to the memory-less property of the exponential distribution (i.e., the model is analytically tractable) (Robbins, T. 2010). Moreover, the modeling of exponential service times is too easy because the exponential distribution has only one parameter (the observed sample mean).

Adopting the Erlang-C model means that servers are assumed to be homogeneous and service-times are independent and identically distributed. The diversity in human-servers' performance, however, is high and tangible because of the varying experience and skills of different agents. In addition, the learning ability of human-servers allows them to be more experienced with time, and thus faster and more efficient. That represents a direct influence on the service times (Armony, M. and Ward, A.2010). Agents suffer also from a shift-fatigue phenomenon where their performance declines during the shift. They perform slower at the end of the shift, especially, if they are kept under stress without short-term relief periods (Sisselman,M. and Whitt,W.2007). Furthermore, it is noticed that the service rates are positively correlated with the arrival rates, that is, the agents perform faster when the offered load is high (Brown, L. et al.2005). Considering such facts in modeling service times brings the validity of the exponential model down.

Quantitatively speaking, the assumption of *Independent-Identically-Distributed* (IID) exponential service times should be validated by looking beyond the sample mean of the data (Green, L. et al.2007). The squared coefficient of variation (SCV) is defined as the squared ratio of the sample mean to its standard deviation. For SCV-values less than one, the Erlang distribution is a good candidate. Finding a parametric model that presents a reasonable fit to the service time is an important issue for the researchers in the call centres field. The Log-normal distribution is sometimes an attractive candidate to model service times in call centres as has been validated empirically in (Mandelbaum, A.2002) and supported in (Brown, L.2005 et al.) and (Pichitlamken, J.et al.2003).

2.3.3 Abandonment Volumes Models

A well-known phenomenon in several service systems is that some of the waiting customers leave, randomly, the queue after waiting for some time. This phenomenon, called customer abandonment, exists in many call centres. The amount of customers that abandon the queue is considered as a direct loss for the call centre (Green, L. et al. 2007). From the input modeling perspective, the abandonments are represented by the abandonment rates, defined as the percentage of callers who hang up while on hold before talking to an agent. The observed abandonment behavior is, originally, the result of certain conditions such as waiting times, and customer patience but must be represented as an input for decision making tools (Saltzman, R. and Mehrotra, V. 2001). Intuitively speaking, the abandonment rates and waiting times in the tele-queues experience strong correlation. Simply, all customers who abandon the queue have waited for some time. At this point, a fundamental distinction should be made. Each customer needs to wait for some time before being served, this time may be less/greater than the time that a customer is willing to wait before abandoning the system (patience time) (Brown, L. et al.2005).

Although that there is a good body of literature regarding the customers' abandonments in call centres such as (Brown, L. et al. 2005),(Green, L. et al. 2007),(Saltzman, R. and Mehrotra, V.2001),(Zohar, E., Mandelbaum, A.2002) and (Deslauriersa, A.et al. 2007), they are all strongly influenced by the empirical data structure. The collected abandonment data, specifically, may differ significantly from one call centre to another in structure itself and the level of detail. Some of the existing work, for example, models the abandonment behavior just using single patience time allowance and apply that for all customers. Other existing work models that behavior using also the patience time but in addition consider another parameter which is the percentage of customers who are willing to wait for long time periods; that is, customers with infinite patience.

2.4 Simulation Modeling and Analysis of Call Centre Dynamics

Simulation usage, specifically computer simulation, in call centre research has not been as popular as some of the other areas in earlier years but has recently become quite important in call centre research. Simulation usage in call centre research is important because call centres, even of the smallest of size, can be quite intricate and complicated when it comes to its inner workings. This is mostly due to the fact that call centres are using advanced technology such as automatic call distributors (ACDs), interactive voice response (IVRs) and computer telephony integration (CTIs) to help aid in answering incoming calls and/or routing both callers and caller information to available agents. CTI allows information to pass back and forth between the IVR and ACD. With that information, the system can orchestrate a screen pop, the simultaneous delivery of a call to an agent's telephone and a screen of information to the same agent's workstation (Robbins, Medeiros, & Dum, 2006)

2.4.1 Realistic Modeling of the Queuing Experience

As discussed in previous chapter, a call centre service system could be modeled as a typical queuing system. On the process side, in a service queue, customers may experience blockings, delays, or need to abandon or retry to obtain the required service. At the other extreme, they may acquire their anticipated service as soon as they put the phone on their ears. Thus, the queuing experience is considered as a window to the service-providing party, through which customers judge the service provider's performance whether it is going better or worse. On the management side, queues could be utilized as indicators for control and improvement opportunities. Consequently, according to (Koole, G. and Mandelbaum, A.2002), queuing analysis provides unbiased quantifiable measures, in terms of which performance is relatively easy to monitor and goals are naturally formulated. In particular, the essence of call centre operations management is the relationship between queuing behavior and staff level decisions. The key question that appears now is how to carry out performance analysis so as to model the attained service level and resource utilization as functions of the demand or workload and available resources.

2.4.2 Performance Evaluation

Performance could be measured through both qualitative and quantitative analysis. The qualitative part corresponds to the psychological aspects of the provided service such as the satisfaction with the provided service itself, whether the agent is friendly or not, and so on (Aksin, Z. et al.2007). The complementary part, quantitative measures, quantifies the performance in terms of service accessibility such as the queue waiting time, the probability of blocking and the retrials. The social sciences and marketing provide empirical models that handle the qualitative measures of performance (Aksin, Z.et al.2007); this kind of measures is out of the scope of the current work. The focus here is on the quantitative management of call centres. In this approach, the workforce staffing and scheduling decisions are made with the objective of operational cost minimization under queuing performance measures (e.g., service level) constraints. In this regard, both analytical modeling (e.g., queuing theory models) and/or experimental modeling techniques (e.g., discrete event simulation models) could be used to provide the quantitative measures of the queuing system performance. In this study, discrete-event-simulation is used in the performance assessment rather than queuing theory. The justification for this choice is derived from literature as further elaborated in the following section.

2.5 Performance Analysis and Evaluation Techniques

2.5.1 Queuing Theory Models for Call Centres

Call centres dynamics are represented using queuing models where the customers are callers, the servers are human agents and/or *interactive voice response* (IVR) units, and tele-queues are the arena where the callers wait for the service provided by the servers. Traditionally, Erlang-C (M/M/s queue) queuing model is used to model the multi-server industrial systems (Gans, N. et al. 2010). For the call centre application, however, the Erlang-C model is an over-simplification of the real queuing system. That model is based on four main assumptions as described in (Gans, N. et al. 2003) and (Robbins, T. et al.2010). The first two assumptions are: that the call arrivals follow Poisson process, and the service-times are independent and identically distributed according to an exponential distribution. The third assumption is that the customers have infinite patience; hence, they never abandon the queue. Finally, the system has infinite number of trunk lines K where the callers will never be blocked. This clearly contradicts the dynamics of the real-world call centres. Thus, it seems that the sole reason for using the M/M/s queuing model is the existence of closed form expressions for most of its performance measures, such as in (Ertogral, K. and Bamuqabel, B.2008), regardless of its ability to model the system accurately.

There are several extensions and improvements to the M/M/s model which exist in the literature. The use of general distribution for service time, for example, the M/G/s model relaxes the second assumption of the M/M/s model. The M/G/s model, however, is not analytically tractable compared to the original model (M/M/s) and its performance measures are calculated as approximations computed in terms of the M/M/s measures. The average waiting time for the M/G/s model, for example, is the product of a factor $(1+C^2)/2$ and the average waiting time under the M/M/s model, where C is the coefficient of variation of the service time (mean / standard deviation; i.e. E/σ) [20]. Thus, it is assumed that the performance deteriorates/improves as stochastic variability in service times increases/decreases.

The simple call centre queuing model, described previously, depicts that each caller within the call centre seizes a trunk line. When all the trunk lines are occupied, the caller receives a busy signal and is blocked. Erlang-B model, where B refers to the term "blocking", accounts for that fact and is denoted as the M/M/s/s queue (Gans, N.et al. 2003). In that queue, there are a finite number of available slots for the incoming callers; this number equals the number of servers (in these case trunk lines). Thus, blocking occurs when there is an incoming caller, but all the servers are busy. The model applies under the condition that an unsuccessful call is not queued or retried, but instead really lost forever which is not also the real-world case.

A more appropriate framework would be the M/M/s/B + M queue or Erlang-A model (Garnet, O.et al.2002). The Erlang-A model assumes that the number of trunk lines is greater than the number of available servers (B > s) so that the excess in the incoming calls could wait for the service in the remaining buffer area (B – s). The waiting customers also have finite patience time that follows exponential distribution. It seems now that Erlang-A model is close to reality, but it still assumes Poisson arrivals, exponential service and patience times.

Beyond the mentioned basic queuing models in queuing theory, the literature includes a substantial vast body of research work that formulates and analyzes queuing models with a realistic essence such as the work in (Deslauriersa, A.et al. 2007), (Whitt, W.2005) and (Dietz, D.2011). However, by reviewing the current literature, most of the research work is theoretically-driven work. Other research work is application-oriented. However, most of that work does not go far enough in the direction of finding a realistic practical solution to modeling such complicated queuing systems as call centres, and few researchers carry the work as far as validating the model or the solution.

One of key reasons for the gap between the theoretical queuing models and representative empirical models of a call centre as a queuing system seems to be the difficulty of evaluating the service operations in call centres. This matches with the discussion on the input data analysis and modeling problems such as the high level of uncertainty, over-dispersion, and the existing strong correlation between different data points in consecutive time periods. Another crucial reason for this gap, as discussed by Robbins and Medeiros (Robbins, T. and Medeiros, D.2006), is that stationary measures are calculated for each daily period, implicitly assuming the system quickly attains stability and reaches the steady state performance. In real call centres, however, which are subject to highly variable arrival rates may rarely achieve steady state. Meanwhile, the analysis of the transient behavior in queuing models literature is scarce (Robbins, T. and Medeiros, D.2006).

2.5.2 Why Discrete Event Simulation (DES)?

Call centres have become a primary means of communication in the modern business world. The appropriate adaptation of advanced telephonic technologies can enhance the capability and efficiency of a call centre (Cleveland and Mayben,1997).. However, these technologies also lead to new features, such as multi-class arrivals, networked queuing stations, skill-based routing schemes, and cross-trained agents, which complicate the management of call centres and violate the assumptions embodied in traditional Erlang-based calculations for traffic flows and queuing. Discrete-event simulation therefore has emerged as a primary tool for the design and operation of call centres, providing decision makers with richer insights into the decision-making process. Recent advances in simulation related techniques, such as simulation optimization, sensitivity analysis, and ranking and selection of alternative designs, promise to help decision makers solve complex problems. Discrete-event simulation models are capable of capturing operational and managerial processes and complex system architectures, in terms of organization, routing policy, and staffing strategy, and can play a vital role in assisting call centre management (Chokshi, 1999).

According to Bouzada (Bouzada, M.2009), the discrete event simulation models usually fit better to the management needs. Although the simulation results are not as precise as the theoretical ones obtained by analytical methods (i.e., queuing theory models), they are often close to them. The precept of the simulation could be stated as "it is better to have rough performance estimates for a very realistic model than exact estimates for a model with several approximations leading to over-simplification". In particular, all queuing theory assumptions described in the previous section are replaced by the facts of real-world systems in the simulation models.

As studied in (L'Ecuyer, P.2006), (Robbins, T. and Medeiros, D.2006) and (Mehrotra, V. and Fama, J.2003), comparing queuing theory and DES models as modeling techniques showed that DES models are superior in several points of the comparison. DES models, for example, cope with more than one type of calls. They also are able to model proactive calls that appear in applications such as telemarketing, charging calls...etc, or as a return for a received call. The high level of uncertainty in input data models involving doubly stochastic models could not be accounted for in queuing formulae. Those models are used, easily, to run random number generators in DES models. The routing of the calls for different agents' groups make the logic behind the call centre even more sophisticated and could not be represented using queuing formulae. DES models succeed in modeling call blockings, abandonments, retrials, general service and patience durations, and general arrival patterns. Thus, the simulation enlarges the capacity of the analytical tools and constitutes a superior approach when there is no theoretical model capable of providing a reasonable system representation to be able to measure the system's performance.

2.6 Optimisation of Staffing and Scheduling Decisions

2.6.1 The Staffing Problem

Determining the optimal required staff levels in the different daily periods is a critical decision in the call centre context. This is due to the direct impact of any minor variation in staff levels on the output performance of the call centre. In order to obtain realistic optimal solutions for this problem, the high level of uncertainty in call centre input parameters should be considered. The uncertainty in these measures, especially in the arrival rates (service demand), is the main source of deviations of the performance measures from the predicted values at the moment of planning (Roubos, A. et al. 20011). Typically, staffing formulations seek to determine the number of full-time equivalent employees needed to achieve cost minimization objective under service level constraints (L'Ecuyer, P.2006).

Traditionally, call centre planners ignore (or model very crudely) the uncertainties and assume known and stationary mean values for the input parameters, mainly for the purpose of tractability. Classical queuing model such as Erlang-C (M/M/c) is commonly used to estimate the stationary system performance in each daily period (e.g., an-hour) (Jongbloed, G. and Koole, G.2001). It is assumed that the average arrival and service rates are constant in each daily period and that the system reaches steady state at the beginning of each daily period. With these assumptions, Erlang-C formula could be used to compute the minimum number of customer service representatives needed to meet a strict service level in each daily period. When that number is large, it is often approximated by the square root safety staffing formula, based on the Halfin-Whitt heavy-traffic regime, and which says roughly that the capacity of the system should be equal to the workload plus some safety staffing which is proportional to the square root of the workload (Gans, N. et al.2003). Erlang-A formula is also used in the same manner in the case of considering customers abandonments (Garnet, O.2002). This approach of obtaining staffing requirements is usually called the Stationary Independent Period by Period (SIPP) approach.

One of the limitations in using the SIPP approach is the dependence on oversimplified assumptions associated with queuing formulae. Another critical problem of the SIPP approach is the assumption of the independence of different planning periods. Pursuant to this, the customers that are waiting in a period will not be carried over to the next period, especially in heavy traffic periods, likely providing understaffed results for the later period (Castillo, I. et al. 2009). As the planning periods get shorter, the stationary state assumption is also questionable in the SIPP approach. The SIPP approach provides stationary measures calculated for each daily period and thus implicitly assumes the system attains stability and reaches the steady state performance. However, due to high variability in arrival rates, the steady state is rarely reached (Robbins, T. and Medeiros, D.2006).

In response to these limitations, different methods have been studied in literature to handle the staffing problem more realistically. One of these methods was done by (Ingolfsson, et al.2007), is based on solving the Chapman-Kolmogorov forward equations by using small, discrete intervals to approximate the continuously varying parameter.. Two other methods are developed by (Whitt, W.2006), one uses a deterministic fluid approximation model and the other uses a numerical algorithm based on a purely Markovian birth-and-death model, having state-dependent death rates. The former method is extended and developed in (Feng, C.2011). All these methods are still dependent on the queuing models but succeed in handling the second and third limitations of SIPP approach.

Call centres could be somewhat complex in structure, and this complexity could make it difficult to identify a queuing model of the centre that is both mathematically tractable and represents a reasonably close match to the real system. In such cases, simulation is a viable alternative. Indeed, simulation is now increasingly used in solving staffing problems. The simulation model allows doing a what-if-analysis to allocate the optimal staffing levels such that some service requirement is satisfied. Moreover, heuristic descent methods could be used to guide the simulation runs in searching for the optimal solution (Mason, A. et al.1998).

2.6.2 Shift Scheduling Problem

Considerable amount of available research were attempting to solve the shift scheduling problem through various approaches with different perspectives (Gans, N. et al.2003). The traditional approach is to solve scheduling problems in two separate steps. After an appropriate staffing level has been determined for each period in the first step, this time-varying staffing is fed into an integer linear programming model to choose, from all feasible shifts covering the staffing requirements, the solution achieving the minimal cost. This approach of finding staff schedules, namely the set covering formulation, was introduced for the first time by Dantzig (Dantzig, G.1996). Several extensions have been introduced to that formulation due to practical purposes. Large call centres that have, for example, several possible shifts and break combinations and associated restrictions require large size mathematical programs with thousands of variables. This issue is addressed by several researchers to provide "implicit formulation" representing shifts and/or breaks by their starting-time windows such as (Aykin, T.1996),(Jacobs, L. and Brusco, M. 1996),(Brusco, M. and Jacobs, L.2000),(Aykin, T. 2000) and (Brusco, M. and Jacobs, L.2001). Those formulations results in a considerable reduction in the size of the mathematical model and thus the computational burden.

The two-stage approach, that solves the staffing and scheduling problems separately in sequence, has its own limitation. The constraints on admissible working shifts often force the second step solution to overstaff in some of the periods. This shortcoming of the two-stage approach has been pointed out in literature with several solution alternatives being proposed. An iterative cutting plane method, for example, is presented in (Atlason, J. et al.2004), to solve a sample average approximation of the problem using simulation and linear programming for minimising staffing costs in a single-skill call centre. The linear integer problem is solved to run the simulation model with the staffing levels obtained from the solution in order to obtain the service level in each period as a function of the staffing level in that period. If the service levels meet the service level constraints then stop with an optimal solution. If a service level constraint is violated, then a new iteration starts with adding a linear constraint to the relaxed problem that eliminates the current solution but does not eliminate any other feasible solutions. Such integrative methods, combining staffing and scheduling problems into a single optimisation problem, could produce more efficient solutions.

2.7 Chapter Conclusion

In view of the large gap between the call-centre reality and queuing models that are tractable, simulation is a key tool for accurate performance estimation and for optimisation. Once the model is well-defined, there is no fundamental difficulty in principle to incorporate its detailed stochastic behavior in a simulation program. However, the accuracy of the used input data models is a key factor in realistic modeling of the call centres. With respect to the most important input parameters, the service time should be modeled as a stochastic input parameter whose values vary from one daily period to another. The most common and suitable statistical model to represent the **service time**, as reported in literature, is the log-normal model.

Regarding the arrival counts model, it needs a really more complicated model than the one used in modeling service times. This is due to the different problems inherent in the arrival counts data which hinder using the traditional models to represent it. It has been concluded that the **arrival counts** need to be modeled as a doubly stochastic process, having a statistical structure for itself (a Poisson model) and another one for its rate. The issue of constructing a suitable statistical model for its rate would be considered further in the following chapter.

The abandonment process is modeled in terms of two different but interrelated parameters. These parameters are the patience time and the percentage of customers that are willing to wait for long time periods. The models for both parameters are also discussed and established in the following chapter.

The realistic input data modeling and simulation help greatly in solving the staffing problem to determine a realistic minimum number of staff members in each of the different daily periods. Traditionally, those realistic estimates are used as inputs to the set covering problem to solve the shift scheduling problem to optimality. The following two chapters demonstrate the methodology adopted in developing and combining the statistical data models, simulation, and optimisation models to achieve the objectives of this thesis.

Next chapter will emphases on the methodology of stochastic input data analysis and modeling as the first step in developing the frame work.

CHAPTER THREE

STOCHASTIC INPUT DATA ANALYSIS AND MODELLING

CHAPTER THREE

STOCHASTIC INPUT DATA ANALYSIS AND MODELLING

3.1 Introduction

The objective of the current research is to develop a framework to enhance the call centre performance through obtaining realistic optimal staffing and scheduling decisions. In order to achieve that, the special features characterising the call centre system should be considered. General characteristics include that call centre systems are complex, most of their operational parameters exhibit highly stochastic nature and require some form of optimisation such as in workforce management decisions. Traditional analytic queuing models and optimisation methods, in contrast, only apply if the systems are sufficiently simple and simplifying assumptions are made. Thus, a combination of simulation and optimisation is more beneficial in developing the proposed performance enhancement framework.

The first step in the proposed framework is to develop appropriate stochastic models for the various operational parameters; namely, arrival counts, service times and the two closely related parameters of abandonment rates and patience times. The next step is to build a simulation model which captures the main characteristics of the call centre dynamics based on the previously developed stochastic models for the various input parameters. The simulation model is in turn used to determine the minimal staffing levels in each daily one-hour period. Finally, after an appropriate staffing level has been determined for each period, this time-varying staffing is fed into an integer linear programming model to choose, from all feasible shifts covering the staffing requirements, the solution achieving the minimal cost.

3.2 The Proposed Input Data Models

In this section, two new models are proposed. The first one is for the arrival counts and the second one is for the abandonment rates. As for the service time model used in the current work; it is extracted directly from the literature to be log-normal.

3.2.1 Daily Proportion-Based Arrival Counts

Recall the doubly stochastic model for arrival counts discussed in the previous chapter. *Gamma-dependent arrival rate* has been used to account for the various difficulties inherent in the nature of the call arrivals data. However, reaching a model which more closely fits the data and the reduction of the required number of parameters to be estimated are major concerns. In the suggested model, the arrival rate in each daily time period is formulated in terms of a proportion P_i of the total daily volume of arrivals Y_j where

$$\Lambda_{ij} = P_i . Y_j, \qquad where\ i = 1, 2, \ldots, k\ and\ j = 1, 2, \ldots, 7 \tag{3-1}$$

where k is the number of daily one-hour periods.

Total Daily Arrivals Model

Proposition 1. The total number of daily calls arriving to a call centre Y_j is normally distributed with probability density function

$$f_Y = \mathrm{N}(\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} . e^{\frac{-(y-\mu)^2}{2\sigma^2}} \quad where - \infty < y < \infty, 0 < \mu < \infty, and\ \sigma > 0 \qquad 3\text{-}2$$

Proof. Given that $Y_j = \sum_{i=1}^{k} X_i$, the proof of *proposition* 1 follows directly from the central limit theorem. ■

The central limit theorem states that the sum of a large number of random variables has a distribution that is approximately normal. It also explains the remarkable fact that the empirical frequency of so many natural populations exhibit bell shaped (that is, normal) curves. The central theorem is extended in (Andrews, D.1991) to be applied to dependent non-identically distributed random variables. This is the case with call centre total daily volume of arrivals which is a sum of dependent, but non-identically, distributed random variables (i.e., call arrivals in each daily period). Hence, Y_j is assumed to be normally distributed according to the central limit theorem. Furthermore, this assumption is supported empirically from the observed data by testing the goodness of fit of the total daily volume data to the normal distribution. Kolmogorov-Smirnov test is used to assess the goodness of fit revealing high p-values which means it yields a very good fit supporting the theoretical assumption of normally distributed total daily call volumes (Mitchell, B.1971).

Due to the high level of variability in total daily call volume, a separate daily arrival normal model needs to be estimated for each day of the week (i.e., there may be seven separate models for total daily arrivals). In order to verify the need to these separate models, the statistical significance of the difference between different samples of the daily call arrivals for different week days is studied. This could be done using the two samples t-test of hypothesis. The hypothesis testing on the difference between the means μ_1 and μ_2 of two normal populations is considered. Suppose that we are interested in testing whether the difference in means $\mu_1 - \mu_2$ is equal to a specified value δ_0. Thus, the null hypothesis will be stated as

$$H_0: \mu_1 - \mu_2 = \delta_0 \qquad 3\text{-}3$$

Obviously, in many cases, $\delta_0 = 0$ is specified to test the equality of two means (i.e., H_0: $\mu_1 = \mu_2$). Suppose that the alternative hypothesis is

$$H_1: \mu_1 - \mu_2 \neq \delta_0 \qquad 3\text{-}4$$

Now, a sample value of $\bar{x}_1 - \bar{x}_2$ that is considerably different from δ_0 is evidence that H_1 is true.

Remark 1. Measuring the significance of statistical difference between different samples of the daily call arrivals for different pairs of week days yields an important result. The pairs that fail to reject the null hypothesis (i.e., those pairs which do not show significant statistical difference) could be dealt as clusters of days having the same total daily arrivals distribution.

This result could be used to reduce the number of estimated parameters and distributions to deal with.

The Proportions Model

As for the proportion value P_i, it is assumed to be independent from the daily volume itself, and thus, it does not differ significantly from day-to-day during the week. This assumption could be verified empirically by showing that there is no significant statistical difference between different samples of the same daily period proportions for different week days. Pursuant to this assumption, the daily proportions of different daily periods could be defined by a single vector **P** = (P_1, P_2... P_k). P_i is the ratio between the arrival rate of period i and the total arrival rate of the day. Then,

$$P_i = \frac{\Lambda_i}{\sum_{i=1}^{k} \Lambda_i} \qquad 3\text{-}5$$

Λ_i is distinct from Λ_{ij} where Λ_i is the arrival rate in period i independent from the week days.

Proposition 2.

i. The daily proportions vector **P** has a *Dirichlet distribution* with parameters (α_1, α_2...α_k) with probability mass function

$$f(p_1, p_2 \dots p_{k-1}) = \mathrm{D}(\alpha_1, \alpha_2 \dots \alpha_k) = \frac{1}{c}.\prod_{i=1}^{k} p_i^{\alpha_i - 1} \qquad 3\text{-}6$$

over the k-dimensional simplex S_k defined by inequalities $P_i > 0$ (i = 1, 2...,$k - 1$), $P_k = 1 - \sum_{i=1}^{k-1} P_i$ and $\sum_{i=1}^{k} P_i = 1$. Here, c (normalisation constant) is the multinomial beta function with the following expression

$$c = \frac{\prod_{i=1}^{k} \Gamma(\alpha_i)}{\Gamma(\alpha_0)}, \qquad \textit{where } \alpha_0 = \sum_{i=1}^{k} \alpha_i \qquad 3\text{-}7$$

ii. The marginal probability distribution of P_i is a *Beta distribution* defined on the interval (0, 1) having parameters (α_i, β_i) with probability mass function

$$f_{P_i} = \begin{cases} \dfrac{\Gamma(\alpha_i + \beta_i)}{\Gamma(\alpha_i).\Gamma(\beta_i)}.p_i^{\alpha_i - 1}.(1 - p_i)^{\beta_i - 1} & 0 < p_i < 1, \quad \textit{where } \beta_i = \alpha_0 - \alpha_i \\ 0 & \textit{otherwise} \end{cases} \qquad 3\text{-}8$$

Assuming that the arrival rates Λ_1, Λ_2..Λ_k of the different daily periods are independent *Gamma random variables* with parameters $\alpha_i > 0$, respectively.

This proves result (i) of *proposition 2*. Result (ii) is a known result derived from the probability mass function of the *Dirichlet distribution*, which is mentioned and proven in (Ferguson, T.1973). Moreover, the assumption that the arrival rates Λ_1, Λ_2...Λ_k are *Gamma random variables* is verified empirically by testing the goodness of fit of the arrival rates to the *Gamma distribution*.

Through the analysis of the observed data, the validity of result (ii) has been recognised to represent a suitable distribution to fit the proportions of daily periods. Applying goodness of fit tests for the *Beta distribution* supports the plausibility of the assumption. The suitability of the *Beta* distribution also follows from the fact that the beta distribution is used in modeling continuous random variables which take on values that lie between 0 and 1, such as proportions and percentages. The fact that the *Beta distribution* is an appropriate distribution for the individual period's proportions motivated the use of its multivariate generalisation, *Dirichlet distribution*, to achieve the correlated tuple of call arrivals.

Based on the above discussion, the number of arrival counts to a call centre at a certain time period i in a given day $j\{X_{ij}\}$ is a *Poisson* random variable with rate Λ_{ij} where the following assumptions hold:

i. $\Lambda_j = (\Lambda_{1j}, \Lambda_{2j},..., \Lambda_{kj})$, where $\Lambda_{ij} = P_i . Y_j$

ii. $\mathbf{P} = (P_1, P_2...P_{k-1})$ where $\mathbf{P} \sim D(\alpha_1, \alpha_2 ... \alpha_k)$ *and* $P_k = 1 - \sum_{i=1}^{k-1} P_i$

iii. $\mathbf{Y} = (Y_1, Y_2...Y_7)$ where $Y_i \sim N(\mu_j, \sigma_j)$

Parameters Estimation

In order to estimate the parameters of the new model, the parameters of both the *Normal distribution* and the *Dirichlet distribution* need to be estimated. Let $[\mathbf{Y_q} = (Y_{1,q}, Y_{2,q}, ..., Y_{7,q})]_{q=1}^{m}$ represent the data set of total daily arrivals observations – where m is the number of weeks in the data set – then both parameters of the normal distribution μ_j and σ_j for each week day could be estimated directly from the observed data set through using the following formulae.

$$\mu_j = \bar{Y}_m = \frac{1}{m}\sum_{q=1}^{m} Y_{j,q}, \qquad \text{where } j = 1, 2, ..., 7 \qquad \text{3-9}$$

$$\sigma_j = S_m^2 = \frac{1}{m-1}\sum_{q=1}^{m}(Y_{j,q} - \mu_j)^2 \qquad \text{3-10}$$

Unlike the *Normal* distribution, the *Dirichlet distribution* is defined with parameters that do not correspond directly to either the mean or variance of the distribution. Rather, the mean and variance of the *Dirichlet distribution* are functions of its parameters α_i. Thus, the parameters of the *Dirichlet distribution* need to be derived using the maximum likelihood estimation (MLE). If the vector $[\mathbf{P_z} = (P_{1,r}, P_{2,r}, ..., P_{k,r})]_{r=1}^{V}$ represents the data set of observed daily period proportions – where V is the number of days in the data set and k is the number of daily periods – then the parameters for a Dirichlet distribution could be estimated by maximising the log-likelihood function of the data (Ronning, G.1989), which is given by:

$$F(\alpha) = log\prod_i \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)}\prod_k p_{ik}^{\alpha_k - 1} = N\left(log\Gamma\left(\sum_k \alpha_k\right) - \sum_k log\Gamma(\alpha_k) + \sum_k(\alpha_k - 1)log\ \bar{p}_k\right) \qquad \text{3-11}$$

where $log\,\bar{p}_k = \frac{1}{N}\sum_i log\,p_{ik}$. Newton-Raphson method is traditionally used to find the unknown parameters.

3.2.2 Random Number Generation

The total daily arrivals are generated using a generator of Normal random numbers. After that, the proportions of each daily period will be generated from the *Dirichlet distribution* using either the *marginal beta distribution* utilizing the result (ii) of *proposition* 2 so that $P_i \sim B(\alpha_i, \beta_i)$ (Ferguson,T.1973)[59]or using the Gamma distribution by finding $Z_i \sim Gamma(\alpha_i, 1)$, then $P_i = Z_i / \sum_{i=1}^{k} Z_i$ and $P_k = 1 - \sum_{i=1}^{k-1} P_i$. The rate of the arrivals in each period is the product of the two generated random variates. Finally, a Poisson arrival count is generated using that estimated rate of arrivals.

3.3 Abandonment Volumes Model

In the work of (Green, L. et al.2007), a customer abandonment phenomenon occurs due to poor service and bad staffing plans; indeed any reduction in abandonment volumes is a direct consequence of improved service and/or staffing plans. Thus, the abandonment volumes should be modeled carefully to create accurate models that represent the operations of a call centre successfully. The modeling of customer abandonment behavior, however, is not an easy task.

The main difficulty here is that the percentage of customers that are willing to abandon and patience time after which they abandon are not absolute input parameters. In essence, their values depend on the system performance (i.e., poor and slow service leads to long queuing time and thus high abandonment volumes). Now, the question is how to model the abandonment behaviour as an input parameter, and at the same time, as an output performance measure. Output performance measures are those with values that are sensitive to the changes in the system design parameters (e.g. number of staffed agents).

In the current work, a new modeling methodology is proposed to account for this difficulty. The proposed methodology represents the abandonment volumes using a couple of parameters, where abandonments are assumed to be of a multi-modal parameter (Saltzman, R. and Mehrotra, V.2001). This assumption is supported by the case study applied in this thesis, where the hazard rate for abandonment (the time phased probability for abandoning) is found to be multi-modal. Therefore, each time-phased mode of abandonment is represented using two parameters. The first parameter is the time interval itself (patience time that after spending it in the queue the customer will abandon). The second is the percentage/weight of customers who belong to this patience-time-interval.

3.3.1 Patience Time Model

The waiting time in the queue till abandonment is divided into n time intervals. Each time interval has a start time $t_s\,(b)$ and end time $t_e\,(b)$; where b = 1, 2... n. The patience time of any customer, considered of patience time allowance within time interval b, is a value falling between $t_s\,(b)$ and $t_e\,(b)$. It is assumed that the all time values within each time interval have equal probability. Thus, the patience time within each interval, $T_{patience}$ (b), is modeled by a uniform distribution.

$$T_{Patience}(b) = UNIF\big(t_s(b), t_e(b)\big) \qquad 3\text{-}12$$

According to this model, the number of time intervals could be infinite. However, the arriving customers are classified into n groups, where n is a limited number (number of time-phased modes/intervals). The group number n is a group which is different from all other groups. It contains customers who are patient enough to wait for long time periods. This group is named *maximum-patience-time-group*. Studying the *maximum-patience-time* shows that it is a stochastic parameter that varies from a daily period (i) to another. The *maximum-patience-time*, T_{max} (i), is assumed to be triangularly distributed with parameters that change from one period to another. Its parameters are defined from the empirical data set of each period to be the minimum, mode, and maximum values of the *maximum-patience-time* recorded values. A single vector is used to define that parameter $\mathbf{T_{max}} = (T_{max}(1), T_{max}(2)\ldots T_{max}(k))$, where

$$T_{max}(i) = TRI\Big(MIN\big(T_{max}(i)\big), MODE\big(T_{max}(i)\big), MAX\big(T_{max}(i)\big)\Big);\ where\ i = 1,2\ldots k \qquad 3\text{-}13$$

3.3.2 Abandonment Group Proportion Model

According to the aforementioned discussion, arriving customers are divided into n abandonment groups. The question to be answered now is regarding the proportion of each group in the total number of arrivals. These proportions also experience strong stochastic nature. Thus, the proportions of different abandonment groups within each separate period could be defined by a single vector $\mathbf{R}(i) = (R_1(i), R_2(i)\ldots R_n(i))$. R_b is the ratio between the arrival rate of customers of group b, $B_b(i)$, and the total arrival rate of the period i. Then,

$$R_b(i) = \frac{B_b(i)}{\sum_{b=1}^{n} B_b(i)};\ where\ i = 1,2,\ldots,k \qquad 3\text{-}14$$

The abandonment group proportions vector $\mathbf{R}(i)$ has a *Dirichlet distribution* with parameters $(\alpha_1(i), \alpha_2(i)\ldots\alpha_n(i))$ with probability mass function

$$f\big(R_1(i), R_2(i)\ldots R_{n-1}(i)\big) = \mathrm{D}\big(\alpha_1(i), \alpha_2(i)\ldots\alpha_n(i)\big) = \frac{1}{a(i)}\cdot\prod_{b=1}^{n} R_b^{\alpha_b(i)-1} \qquad 3\text{-}15$$

over the n-dimensional simplex S_n defined by inequalities $R_b(i) > 0$ $(b = 1, 2\ldots,n-1)$, $R_n(i) = 1 - \sum_{b=1}^{n-1} R_b(i)$ and $\sum_{b=1}^{n} R_b(i) = 1$. Here, $a(i)$ (a normalization constant) is the multinomial beta function with the following expression

$$a = \frac{\prod_{b=1}^{n}\Gamma\big(\alpha_b(i)\big)}{\Gamma\big(\alpha_0(i)\big)},\qquad where\ \alpha_0(i) = \sum\nolimits_{b=1}^{n}\alpha_b(i) \qquad 3\text{-}16$$

The arrivals are considered as a Poisson process. The Poisson process experiences a well-known property named *random selection*. This property states that if a random selection is made from a Poisson process with intensity Λ such that each arrival is selected with probability R, independently of the others, the resulting process is a Poisson process with intensity $R\Lambda$ (i.e., $B_b(i)$).

According to this methodology, the abandonment behaviour is reasonably modeled as will be seen in the chapter five. The abandonments would be sensitive to changes in the system design parameters. Interestingly, if there are improvements in the system performance, then the waiting time of the customers decreases.

3.4 Chapter Conclusion

In this Chapter, a detailed description of the proposed performance enhancement methodology for call centres operations is provided. A new doubly stochastic model has been proposed to represent the arrival counts in different one-hour periods, which successfully addresses the various challenges associated with modeling the arrival process in call centres. Moreover, a new methodology for modeling the abandonment rates has been developed.

CHAPTER FOUR

SIMULATION MODEL AND

OPTIMISATION OF STAFFING AND SCHEULING DECISIONS

CHAPTER FOUR

SIMULATION MODEL AND OPTIMISATION OF STAFFING AND SCHEULING DECISIONS

4.1 Introduction

In this chapter we will create and build a simulation model which captures the main characteristics of the call centre dynamics based on the previously developed stochastic models for the various input parameters. The simulation model is in turn used to determine the minimal staffing levels in each daily one-hour period.

The goal of the optimisation model is to find values of the decision variables that will optimise (maximise or minimise) the objective function among the set of all values for the decision variables that satisfy the given constraints. (Winston, W. L.2003), and (Venkataramanan, M. 2003).

Call centres are becoming increasingly popular and important. Whether it is to deal with customers of a product or service, or to provide a specialists' help to different businesses, everyone deals with call centres from time to time. With this increase in call-centre popularity comes an increase in interest in operations research analysts to design and implement the best combination between staffing and schedules for these centres. Being able to reach a technician or call operator requires a lot of thought and planning, including forecasting number of calls at different times of the day, determining how many agents need to be working to handle these forecasted number of calls, and lastly, how to assign agents to these shifts so that enough people are working to cover the forecast. In small sizes, this problem can be fairly trivial, but once the requirements for workers decreases in flexibility and more and more agents are available, the harder this problem becomes to solve.

4.1.1 Methodology of Conducting a Simulation Study

Use of a computer simulation model to mimic call centres dynamics, as a performance evaluation tool, is a surrogate for evaluation using queuing formulae which usually involve approximations and oversimplified as well as inaccurate assumptions. Yet, if the simulation model is not a close-approximation to the actual system, any conclusions derived from the model would be likely to be erroneous and may result in costly decisions being made.

Having a formal approach for conducting a simulation study is crucial to develop a reliable and valid model. Figure 4-1 presents an eight-step approach, adapted from (Banks 2000), (Law, A.,2005) and (Tako, A. and Robinson, S. 2010), to conduct such a robust simulation study. The basic steps in the simulation process are described in the following sections.

4.1.2 Problem Formulation

The first step in building a simulation model is to analyze the problem itself that the simulation model has been built to address, where the system modeling is rarely undertaken for its own sake. Rather, modeling is prompted by some system-oriented problem (e.g. performance evaluation of the service system of a call centre) whose solution utilizes the developed model. At this stage, the behaviour of the system of interest (which could be a natural or artificial system, existing or not) is studied and analyzed in order to organize the system's operation as objects and activities within the experimental framework of interest. Furthermore, the input/output variables must be identified and classified into decision variables (controllable) or parameters (uncontrollable).
If the problem involves performance analysis, this is the point at which the performance metrics could be identified and defined based on the output variables, in addition to an objective function (i.e., a combination of some of the metrics). In conclusion, the problem formulation is to study and analyze the system behavior, indentify and classify the operations variables, and finally define the performance metrics along with the desired objective function of the study.

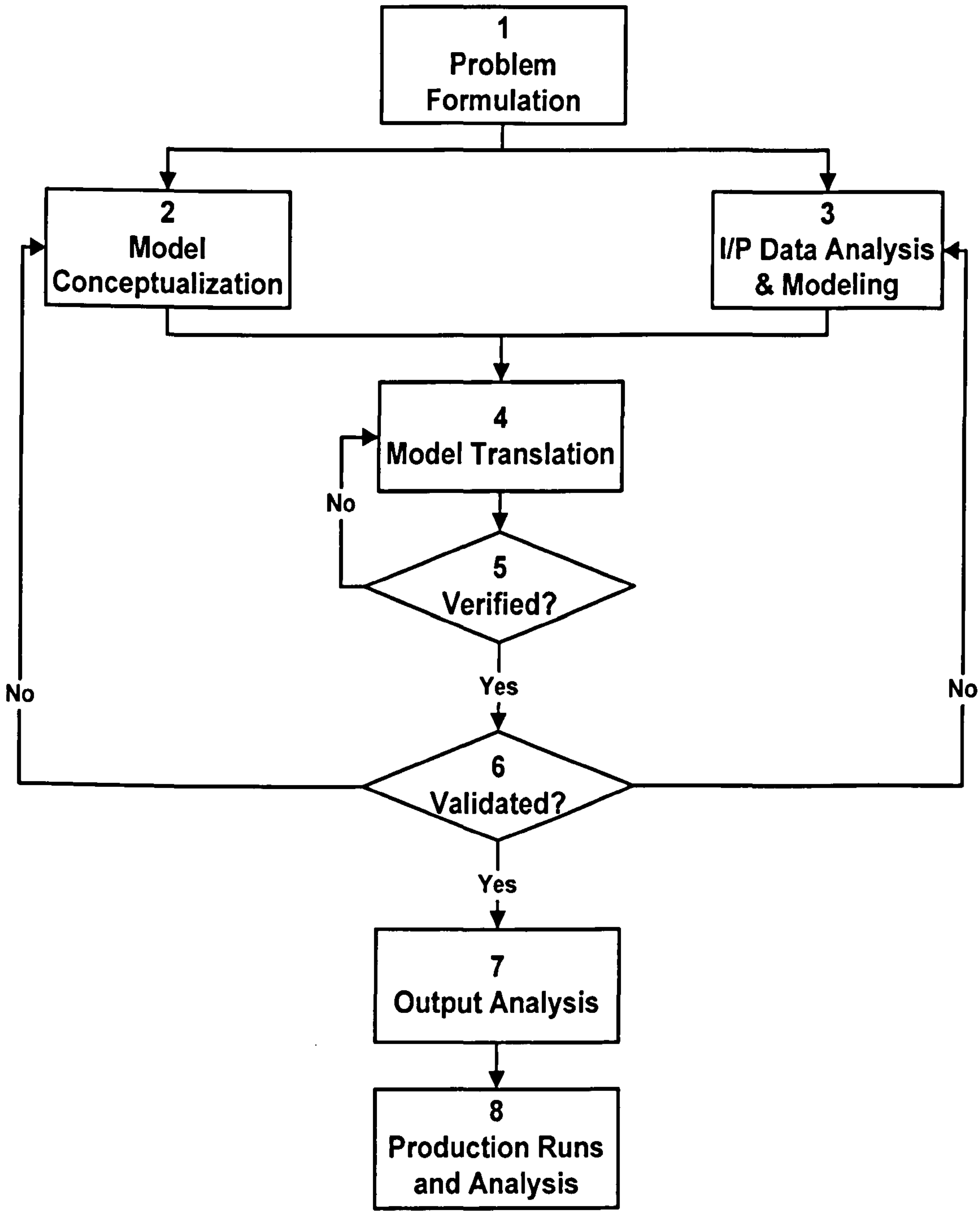


Figure 4-1: Eight-step approach of conducting a simulation study (Tako, A. and Robinson, S. 2010)

4.1.3 Model Conceptualisation

The real-world system under investigation is abstracted by a *conceptual model.* This step consists of building a high-level description of the structure and behaviour of the system and identifying all the objects with their attributes and interfaces. This description is then represented using logic flow diagrams, hierarchy trees, narrative, or any other convenient means of representation. Finally, nonfunctional information must also be documented, for instance, possible future changes, unintuitive (or informal) behaviour, and the relation with the environment.

4.1.4 Input Data Collection, Analysis, and Modeling

In this phase, the attributes chosen in the previous step must be observed and their values recorded. When the system entities are studied, they need to be associated with a time factor. Another important issue during this phase is the selection of a sample size that is statistically reasonable and a data format that could be processed with a computer. Finally, the selected attributes must be classified as stochastic or deterministic attributes. The gathered input data are used to develop models for the stochastic attributes, which could be descriptive, explanatory, and/or theoretical models as described in the previous chapter.

4.1.5 Model Translation

In the modeling phase, once the problem is fully studied (i.e., the conceptual model is built) and the input data is collected and analyzed, a computer simulation model could be constructed and implemented. A simulation model must be ultimately transcribed into computer code, using some programming language. A simulation language may be general purpose or special purpose. A general-purpose programming language, such as C++ or Visual Basic, provides no built-in simulation objects (such as a simulation clock or event list), and no simulation services (e.g., no clock updating or scheduling). In contrast, a special-purpose simulation language (e.g., Arena, Promodel, GPSS) implements a certain simulation worldview, and therefore does provide the corresponding simulation objects and services as built-in constructs. It is strongly recommended that the modeler give considerable thought to the selection of an appropriate simulation language. The main selection criterion is the degree to which the simulation language's worldview fits the type of simulation model to be coded.

In order to translate the model logic of the studied call centre, in this thesis, into a computerised model, the software package *Arena®* provided by Rockwell Automation is used. *Arena®* is a special-purpose simulation software package designed to imitate the system operations or characteristics over time (Kelton, W. et al.2010). This software package uses an entity-based, flowcharting methodology for modeling dynamic processes. It was used because of numerous benefits, as described in (Nagarajan, K. et al. 2003), such as:

1. ***High Usability.*** User friendly and even non-technical people could use it.
2. ***High Understandability.*** Excellent graphical interface for the simulation model, which gives us a clear outlook of how the simulation model works.

3. ***High Flexibility.*** Modeling from simple to complex systems is possible using *Arena.*

4. ***High Functionality.*** Integrity of the controls as well as functional completeness could be achieved using *Arena.*

5. ***High Extendibility.*** *Arena* capacity is so vast and could be extended to a variety of applications.

6. ***High Reliability.*** The simulation results are accurate in most cases as reported by its users, though few discrepancies may occur.

4.1.6 Model Verification

During the previous steps, two different models are built: the conceptual model (specification) and the operational simulation model (executable computer program). Verification is related to the internal consistency among the two models (is the conceptual model accurately represented by the operational model?). Differently stated, verification makes sure that the model conforms to its specification and does what it is supposed to do. Model verification is conducted largely by inspection, and consists of comparing model code to model specification. Any discrepancies found are reconciled by modifying either the code or the specification. Model verification could be conducted using several techniques (applying some of them is satisfactory) such as (Banks 2000):

1. The model should be checked by someone other than its developer.
2. A flow diagram which includes each logically possible action that a system could make when an event occurs should be constructed. The model for each action of each event type should be traced.
3. Closely examine the model output for soundness under a variety of settings for the input parameters.
4. If the model is animated, verify that what is seen in the animation imitates the actual system.
5. The simulation could be monitored as it progresses. This could be accomplished by advancing the simulation until a desired time is elapsed then displaying the model information at that time or advance the simulation until a particular condition is in effect and then display the information.

4.1.7 Model Validation

Although verification is necessary, it is not sufficient (i.e., a model may be verified but not valid). Every verified model should be initially viewed as a mere proposal, subject to validation. Validation focuses on the correspondence between the model and reality (are the simulation results consistent with the system being analysed?). In essence, model validation examines the fit of the model to empirical data (measurements of the real-life system to be modeled). A good model fit means here that a set of important performance measures, predicted by the model, match or agree reasonably with their observed counterparts in the real-life system. Any significant discrepancies would suggest that the proposed model is inadequate for project purposes and those modifications in the *conceptual model* and the *input* *data models* are called for. In practice, the validation process does not constitute a particular phase of the life cycle, but it is an integral part of it. It is common to go through multiple cycles of model construction, verification, validation, and modification.

Validation could be performed through a series of tests, some of them are subjective and others are objective. According to (Banks 2000), ***subjective tests*** usually involve people who know some or many aspects of the system, making judgments about the model and its outputs considering only the practical significance of the obtained results such as in (Bekker, J. and Viviers, L.2008) and (Longo, F.2010). One other technique that is adopted here in this thesis is the *data-driven model validation procedure.* This test uses the recorded data from the actual system as it is instead of creating them by sampling from a probability distribution to develop a deterministic system in which no randomness is involved. If the results of a simulation run that's driven by the recorded data from the actual system closely correspond to the actual system performance over that period of time, then there is greater confidence in the correctness of the conceptual model. Thus, deterministic tests are used when objective tests fail to prove the model validity. This is to determine whether the absence of the model validity is due to problems in the conceptual model, input data models, or both of them.

4.1.8 Output Analysis

Each run of the simulation program, called a replication in simulation parlance, is an experiment that yields a sample system record from which various statistics could be estimated. Output analysis is the stage concerned with designing replications, computing statistics from them and presenting them in either textual or graphical format. Thus, as its name suggests, output analysis focuses on the analysis of simulation results. It provides the main value-added of the simulation enterprise by trying to understand the system behavior and generate predictions for it. The main two issues addressed in output analysis are:

1. ***Replication Design.*** A good design of simulation replications allows the analyst to obtain the most important statistical information from the simulation runs with the least computational cost. In particular, we seek to minimize the number of replications and their length, and still obtain reliable statistics.

2. ***Estimation of Performance Metrics.*** Statistics gathered in various replications provide the data for computing point estimates and confidence intervals for system parameters of interest. Critical estimation issues are the size of the sample to be collected and the independence of observations used to compute statistics, particularly confidence intervals.

4.1.9 Simulation Experiments Time Frame

Simulation models are usually classified into two main classes based on their time horizon (Sargent, R.2010): terminating models and steady-state models. Each class gives rise to different statistical issues relating to their output analysis. *A terminating simulation* is one in which the model dictates specific starting and stopping conditions as a natural reflection of how the target system actually operates. As the name suggests, the simulation will terminate according to some model-specified rule or condition. In terminating simulation models, the number of replications is the critical design parameter of the associated experiment output analysis, since it is the only means of controlling the sample size of any given estimator. In essence, the sample size affects directly the estimator's variance, and consequently, its statistical accuracy.

A steady-state simulation, on the other hand, is one in which the quantities to be estimated are defined in the long run; that is, over a theoretically infinite time frame. In particular, a steady-state simulation model has no natural termination time for its replications, and could be potentially run forever. Thus, only long-term statistics are of interest, but initial system conditions tend to bias its long-term statistics. Therefore, it makes sense to start statistics collection after an initial period of system warm-up, namely, after the biasing effect of the initial conditions becomes insignificant. Consequently, steady-state models have a number of important issues associated with their output analysis; mainly the replication length and the warm-up period. The guiding principle of the replication length is stabilization of the statistics of interest. The length of a warm-up period is determined by observing experimentally when the time variability of the statistics of interest almost disappears.

The Adopted Experiments Time Frame

Deciding on the time-based suitable classification for a twenty-four-hour, seven-days-a- week (24/7) service call centre may seem pretty clear and the decision would be directly a "steady-state" model. The state of almost 24/7 call-centres, however, is highly dynamic, and over time a stochastic system evolves and changes continually. Moreover, transient-time behavior related performance metrics and steady-state performance metrics need to be gathered.

Accordingly, if the objective is to decide on decision variables that are functions of short-term performance metrics, such as staff level of a specific daily period, the call centre system would be treated as a terminating system. In essence, short-term performance metrics, which are almost calculated over an/half hour period, are always in transient state. This is due to the stochastic nature of the service demand (i.e., arrival volumes) which yields the short-term performance metrics not to stabilize at all.

On the other hand, if the focus is on decisions that are function of long-term performance metrics, such as the agents' weekly schedule, the decision is pretty clear to analyze the system as a "steady-state" system. In particular, the agent's weekly schedules are built to satisfy a certain overall performance level over the course of each day and thus over the week period.

4.1.10 Output Data Collection and Analysis of Terminating Systems

As mentioned previously, each replication produces a random record (sample path), from which various statistics are computed. These statistics are estimates of various parameters of interest (probabilities, means, variances, etc.). More formally, if the performance parameter θ is of interest, then the simulation will then produce an estimator,$\hat{\Theta}$, for the true but unknown parameter, θ, which evaluates to some estimate $\hat{\Theta} = \hat{\theta}$.

In essence, statistical output analysis has two general purposes:

1. ***Estimation of the performance measure by a single number***, called the point estimate. The sample mean is classified as a point estimator, because it estimates a scalar. Thus, the used point estimate is:

$$\bar{\Theta} = \frac{1}{n}\sum_{r=1}^{n} \hat{\Theta}(r) \qquad 4\text{-}1$$

2. ***Estimation of the error in the point estimate***, in the form of a confidence interval. In order to quantify the confidence in the point estimate value $\bar{\Theta}$, the probability of events of the form $Pr\{\hat{\Theta}_1 \leq \theta \leq \hat{\Theta}_2\} = 1 - \alpha$ is computed. The estimators $\hat{\Theta}_1$ and $\hat{\Theta}_2$ define a (random) confidence interval $[\hat{\Theta}_1, \hat{\Theta}_2]$ for θ, and α is the probability that the confidence interval does not include θ (α is a small probability, typically around 0.05). Thus, the confidence interval could be obtained in the form of (assuming that $\bar{\Theta}$ is normally distributed)

$$\bar{\Theta} \pm z_{1-\alpha/2}\sqrt{\frac{\sigma^2}{n}} \qquad 4\text{-}2$$

   Equation 4-2 highlights the fact that the confidence interval for the mean value, considered here, is symmetric (two-sided) about the (unbiased) point estimator, $\bar{\Theta}$, with half-width $z_{1-\alpha/2}\sqrt{\frac{\sigma^2}{n}}$. Since the quantile $z_{1-\alpha/2}$ and the variance σ^2 are fixed values, the accuracy of a confidence interval could be enhanced only by increasing the number of replications n.

Remark 4. There is no standard method to determine the suitable number of replications that ensures a reasonably small confidence interval (high accuracy level). An approximation, however, is proposed in (Kelton, W. and et al. 2010). This approximation depends on conducting a simulation run with a starting number of replications n_o that produces an initial half width h_o. After that, the number of replications n to obtain the desired half width h is

$$n \cong n_o \frac{h_o^2}{h^2} \qquad 4\text{-}3$$

4.1.11 Output Data Collection and Analysis of Steady-State Systems

There are several strategies to perform output data analysis for steady-state systems such as: truncated replications, batch means, econometric time-series modeling, standardised time series, regenerative models, as well as variations on batch means like separating or weighting the batches. In this thesis, the method of truncated replications would be suitable for the studied system.

Truncated replications(Kelton, W. and et al. 2010) strategy treats the output data in a similar manner of that of the terminating systems with a small variation. A warm-up period should be identified and be reasonably short relative to the run length. With this appropriate warm-up and run length specified, statistical analysis is carried out as in the terminating systems case with warmed-up independent replications so that the steady-state parameters are obtained rather than terminating quantities. In order to find a suitable warm-up period, the model is run and one of the overall performance measures would be plotted as a function of the simulation run length. The period after which these measures achieve steady-state performance would be a suitable warm-up period.

4.1.12 Production runs and analysis

Statistical estimates are used in turn to understand system behavior and generate performance predictions under various scenarios, such as different input parameters (parametric analysis), scenarios of operation, and so on. Experimentation with alternative system designs could elucidate their relative merits and highlight design trade-offs.

4.2 The Proposed Validation Framework

The main technique used in this section is the objective test supported by the proposed deterministic test whenever needed. In performing the validation process, a complete validation methodology is proposed through the analysis of related literature. This methodology is depicted in flow-chart shown in Figure 4-2.

1. ***Select Validation Parameter.*** The input data of any simulation model could be categorized as controllable, which are decision variables D, and uncontrollable parameters, which are referred to as X. Those decision variables are the input data that could be adjusted to optimise the system performance. After building the model and validating its logic and assumptions, the model are ready to be operated and its results be validated. The resulting model is now viewed as a "black box" which receives all input variables specifications (X and D) and transforms them into a set of output or response variables denoted by Y_n (n is the number of different output variables).

$$Y_n = f(X, D) \qquad 4\text{-}4$$

The output variables Y_n constitute the statistics of interest generated by the simulation about the model's behaviour such as: queues waiting time, service level, abandonment volumes...etc. The choice of the validation output parameter(s) is directly related to the objectives of the project. Here, the simulation output variables are denoted by Y_n, while the real-system output variables are denoted by Z_n.

2. ***Check Normality of Output Populations.*** The normality of output data sets is checked using *Anderson-Darling normality test.* If the obtained P-value is greater than 0.05, then the data sets is considered normally distributed.

3. ***Check Equality of Normal Populations Variances.*** The standard two-sample t-test for testing the equality of the means of two normal populations assumes that the compared data sets have equal variances. However, there are situations in practice where such assumption represents significant inaccuracy. Thus, if the sample variances of real and simulated data are equal then the standard two-sample t-test would be used. Otherwise, the Smith-Satterthwaite method is used for cases in which data show different variances. This test considers the differences in variances by adjusting the degrees of freedom for a critical t-value. In order to check the equality of the compared population's variances, the F-test is used.

4. ***Check Equality of Population Means.*** This step is done through comparing the difference between the model output variables Y_n and the real system results Z_n to a specific amount of allowable difference δ_o (usually small enough to consider it zero). Formally, a statistical test of the null hypothesis is:

$$H_o: E(Y_n) - E(Z_n) = \delta_o$$

versus

$$H_1: E(Y_n) - E(Z_n) \neq \delta_o \qquad 4\text{-}5$$

where H_o is the hypothesis that the expectation of the difference between the performance measure and the simulated one is equal to δ_o, which is the mean of the system parameter Z_n. H_1 is the hypothesis that this difference is not equal to δ_o. In essence, if H_0 is not rejected, then on the basis of this test there is no reason to consider the model is invalid. If H_0 is rejected, the current version model is rejected and the modeler should seek ways to improve the model.

The standard test to perform such statistical comparisons is the two-sample t-test. However, a key assumption of t-test is the normality of tested data sets which may not be the real case. Otherwise, when the distribution under study is not characterized as normal, according to (Fay, M. and Proschan, M.2010), non-parametric methods become very useful such as a *Wilcoxon-Mann-Whitney test.* No assumption about the population distribution in nonparametric methods is made. In the *Wilcoxon-Mann-Whitney test,* given two independent samples, the null hypothesis is tested to see whether the median of two samples are equal.

5. ***Check the Deterministic Validity (data-driven model validation).*** The way to improve the invalid model is through revising and modifying the conceptual and/or input data model. At this point, checking the deterministic validity of the model, as discussed previously, helps to make the distinction of whether the validity defect is due to defects in the conceptual model, input data models, or both of them. If the model is deterministically valid, then the defect is in the input data models. In essence, deterministic validation ensures the correctness of the conceptual model.

The proposed framework will be validated through the proposed validation methodology as shown in figure 4-2

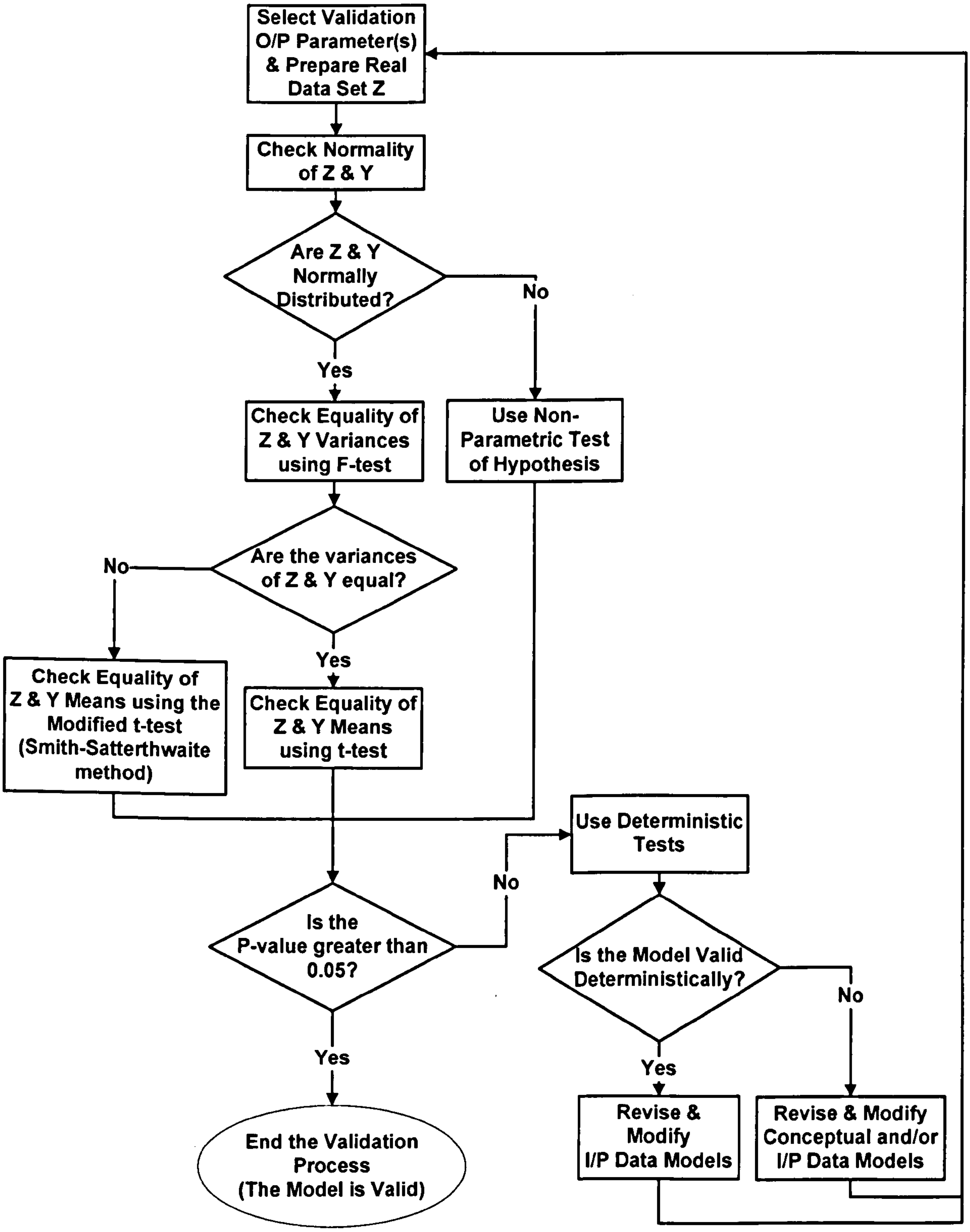


Figure 4-2: Proposed Validation Methodology

4.3 Optimisation of Staffing and Scheduling Decisions

4.3.1 Problem Definition

The staffing and shift scheduling decisions seek to determine an optimal collection of shifts to be worked, such that staff costs are minimised while achieving specific service levels or other labor requirements. A shift is defined as a set of time-intervals in a day during which a customer representative works, and a schedule here refers to a set of shifts that provides the total staffing requirement for a day. Closely related to the scheduling problem, the rostering problem which combines shifts into rosters and provides the actual matching between employees and rosters (Aksin, Z. et al. 2007). The focus of the current work is only the problem of staffing and shift scheduling.

In the staffing and shift scheduling problem in call centres, there are k daily periods, and s different working shifts for the day. The cost vector is $\mathbf{C} = (c_1, c_2 \ldots c_s)$, where c_j is the cost of an agent having shift j. The decision variables vector is $\mathbf{X} = (x_1, x_2 \ldots, x_s)$, where x_j the number of agents having shift j. The staffing vector is $\mathbf{Y} = (y_1, y_2 \ldots y_k)$ where y_i is the required number of agents in period or interval i. This vector $\mathbf{Y}$ could be written as $\mathbf{Y} = \mathbf{AX}$ where $\mathbf{A} = [a_{ij}]$, where

$$a_{ij} \begin{cases} 1, & \text{if an agent working according to schedule j is available to take calls during inte} \\ 0, & Otherwise \end{cases} \qquad 4\text{-}6$$

The service level for period i is

$$g_i(\mathrm{Y}) = \frac{E[No.of\ calls\ answered\ within\ \tau_i]}{E[No.of\ calls\ arriving\ in\ period\ i]} \qquad 4\text{-}7$$

for some constant τ_i where the expectations are with respect to the probability distributions that define the stochastic model (arrival rate process, service times, abandonments...etc.). The corresponding minimal service level for some period j is l_j.

4.3.2 Problem Formulation

In this thesis, as discussed above, a two-stage approach is used to solve the problems of staffing and scheduling decisions. This means that the output of the first problem, P1, is fed as an input for the second one P2. The staffing optimisation problem with service level constraint, solved for each daily period *i*, could then be formulated as

(P1)
$$Minimize\ y_i \qquad 4\text{-}8$$
$$Subject\ to\ g_i(\mathbf{Y}) \geq l_i \qquad 4\text{-}9$$

The shift scheduling problem is formulated as a set-covering *integer program* which has the following formulation

(P2)
$$Minimize \sum_{j=1}^{s} c_j x_j \qquad 4\text{-}10$$
$$Subject\ to \sum_{j=1}^{s} a_{ij}.x_j \geq y_i \quad for\ all\ i = 1, 2, \dots, k \qquad 4\text{-}11$$
$$x_j \geq 0\ and\ x_j\ is\ integer \quad for\ all\ j = 1, 2, \dots, s \qquad 4\text{-}12$$

4.3.3 Solution Procedure for Solving the Staffing Problem

In order to solve the staffing problem, an approximate or estimate for the service level functions $g_i(\mathbf{Y})$ is needed. Each of these functions, in general, may depend on several components of the vector **Y** in a complicated way, because changing the number of agents in one period may change the queue length in the periods that follow, for example. Furthermore, it may change the waiting time of a call that arrived in the previous period. As previously discussed in chapter 2, simplified queuing models may sometimes provide reasonable rough-cut approximations to these SL functions, but simulation seems to be the only way of getting reliable estimates of their values for call centres.

Solving this problem in the context of multi-skill call centres requires a simulation-based optimisation algorithm to determine the optimal staff level in each daily period for each skill group. In a single-skill context, however, it is fair enough to apply some sort of what-if-analysis to allocate the optimal staffing levels such that some service requirement is satisfied as shown by (Pichitlamken, J. et al.2003).

In the current work, a realistic simulation model is developed using a simulation package *Rockwell Arena*®. This model is run several times, for each daily period in sequence, with different staff levels till reaching the staff level that satisfies the service level constraint. In order to avoid the drawback of running each period independently, all previous periods before the evaluated period are considered in the run. However, the performance measures are estimated only for the period under evaluation. Therefore, each period is affected by the call centre performance in the previous periods.

4.3.4 Solution Procedure for Solving the Shift Scheduling Problem

In the current work, LINGO® 12.0 optimisation package produced by LINDO® Ltd. is used to produce an exact optimum solution for the scheduling problem. LINGO® is a comprehensive software package designed for building and solving mathematical optimisation models. It provides a completely integrated package that includes a language for expressing optimisation models, a full-featured environment for building and editing problems, and a set of fast built-in solvers capable of efficiently solving most classes of optimisation models.

For models with general and binary integer restrictions, such as the current scheduling model, LINGO® includes an IP solver that works in conjunction with the linear, nonlinear, and quadratic programming solvers.

A detailed description of the LINGO® implementation of the scheduling model is provided in Appendix A.

4.4 Chapter Conclusion

In this chapter, a detailed description of the proposed performance enhancement methodology for call centres operations is provided. Furthermore, the adopted simulation model validation procedure has been detailed. Finally, a simple simulation-based what-if-analysis procedure is used to determine the staffing level in each one-hour period that satisfies a given service level constraint, which in turn is fed as input to an IP model that solves the shift scheduling problem.

The realistic input data modeling and simulation help greatly in solving the staffing problem to determine a realistic minimum number of staff members in each of the different daily periods. Traditionally, those realistic estimates are used as inputs to the set covering problem to solve the shift scheduling problem to optimality.

In the following chapter, the proposed methodology is applied to a real, single-skill, inbound call centre as a case study demonstrating its validity and potentials for improving call centre performance.

CHAPTER FIVE

APPLYING THE PROPOSED ENHANCEMENT FRAMEWORK TO A CALL CENTRE

CHAPTER FIVE

APPLYING THE PROPOSED ENHANCEMENT FRAMEWORK TO A REAL CALL CENTRE

5.1 Introduction

The studied Call Centre is a modest sized company of approximately 150 agents.

A significant portion of the services provided by this company come from their customer service call centre, who provide assistance to their customers via telephone.

The system under study is the telephone directory call centre of the public service telephone network (PSTN) company in Libya. This call centre is a single-skill inbound call centre. The call centre provides twenty-four hours a day/seven days a week support to its customers. The available data is aggregated over sixty-minute time periods. That is, the twenty-four hours are partitioned into twenty-four periods in each of which the model parameters remain constant and then change when the new period starts for each of the seven week days. A sample of the provided data sheets is presented in appendix B. The structure of the studied call centre and its operations is demonstrated later. In this chapter, the proposed performance enhancement framework is applied to this call centre.

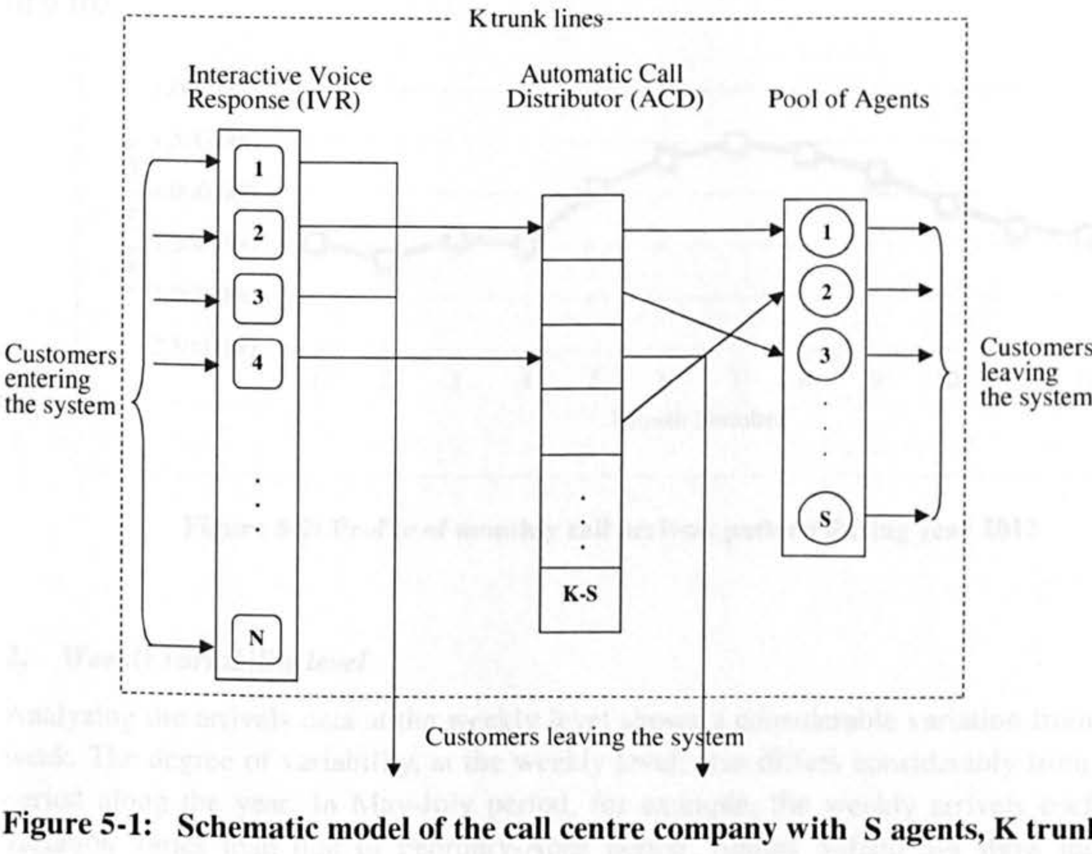


Figure 5-1: Schematic model of the call centre company with S agents, K trunk lines.

5.2 Arrival Counts Model Implementation

5.2.1 Prelude Data Analysis

The analysis of the call centre arrival process reveals the existence of the mentioned three main properties of arrivals data: high level of uncertainty, over-dispersion, and strong correlation.

High Uncertainty Property

The uncertain arrival pattern according to 2.3.1 is a key property in the call centre arrivals and presents several problems in modeling call centres, as discussed in previous chapters. Considering this issue, we find different levels of variability: monthly, weekly, daily, and periodically within the same day.

1. *Monthly variability level*

Considering the monthly arrival volumes in year 2012 reveals that the arrivals pattern is nearly constant over the six-month period starting from January to April and then from November till December. In the other six-month period, the arrival pattern experiences a considerable increase in May continuing till July, and then decreases again. Thus, there is a high level of month-to-month variations in call volumes in the May-October period.

Figure 5-2 represents the monthly arrival pattern in year 2012 with a mean of 3,885,064 arriving calls per month, a standard deviation of 396,789 calls, and a coefficient of variation of 0.102.

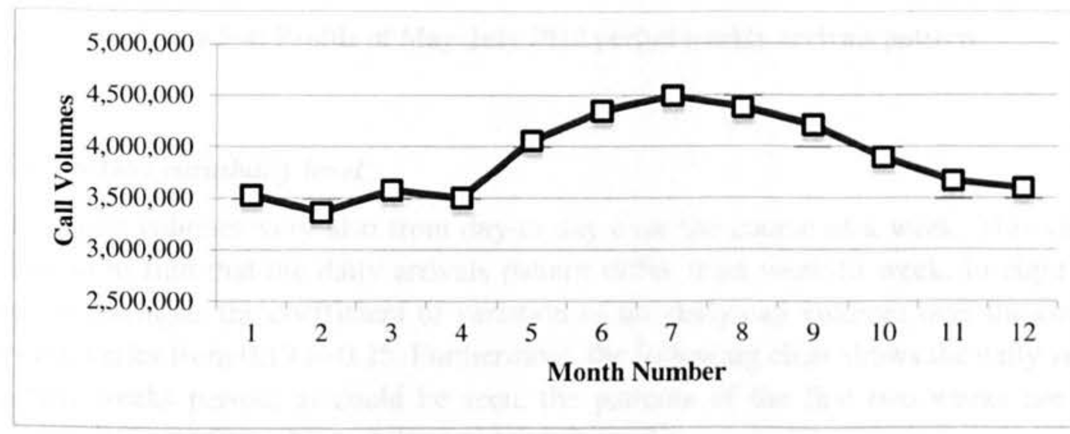


Figure 5-2: Profile of monthly call arrivals pattern during year 2012

2. *Weekly variability level*

Analyzing the arrivals data at the weekly level shows a considerable variation from week-to-week. The degree of variability, at the weekly level, also differs considerably from period to period along the year. In May-July period, for example, the weekly arrivals coefficient of variation varies than that of February-April period; figures 5-3and 5-4 show the weekly-arrivals pattern for both periods.

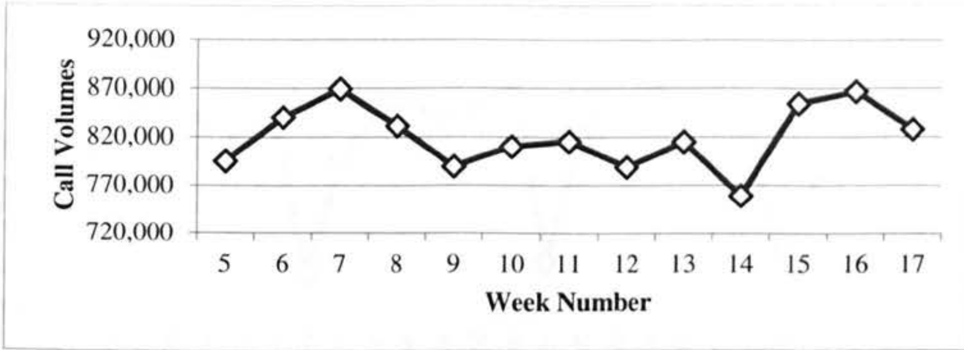


Figure 5-3: Profile of February-April 2012 period weekly arrivals pattern

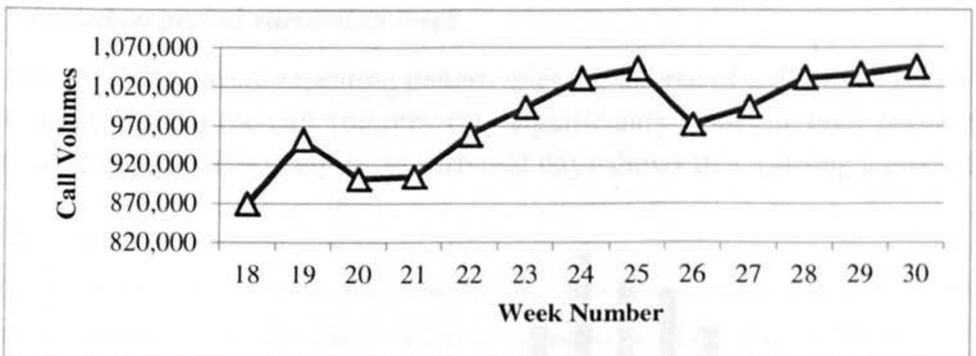


Figure 5-4: Profile of May-July 2012 period weekly arrivals pattern

3. *Day-to-Day variability level*

The daily call volumes vary also from day-to-day over the course of a week. This variation may extend to find that the daily arrivals pattern differ from week to week. In eight weeks period, for example, the coefficient of variation of the daily call volumes over the course of each week varies from 0.19 to 0.25. Furthermore, the following chart shows the daily volumes along four weeks period; as could be seen, the patterns of the first two weeks are nearly similar but different from those of third and fourth weeks.

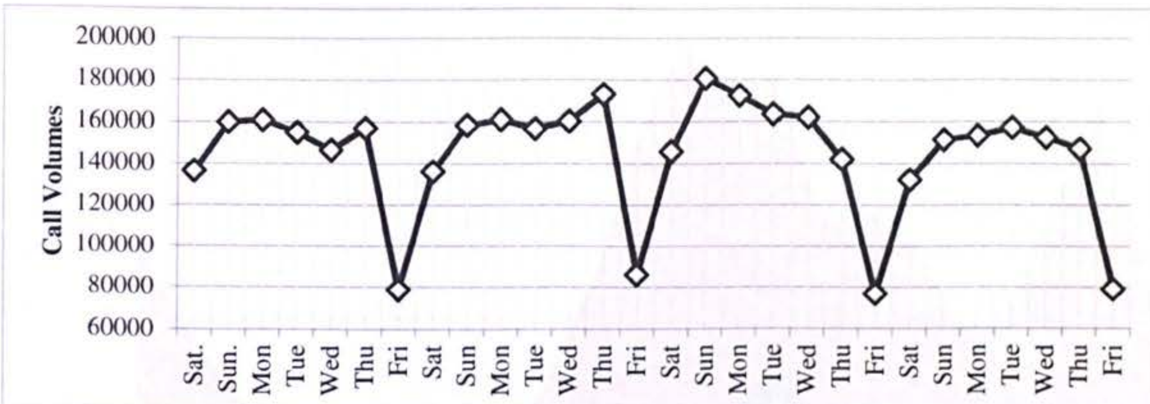


Figure 5-5: Profile of daily arrivals during first four weeks of June 2012

4. ***Daily period-to-period variability level***

Call volumes exhibit a strong repeating pattern over the course of a day. The following chart, Figure 5-6, indicates that the call volumes vary significantly from one hour-period to another during the same day. Similar analysis for different days shows that a strong pattern exists.

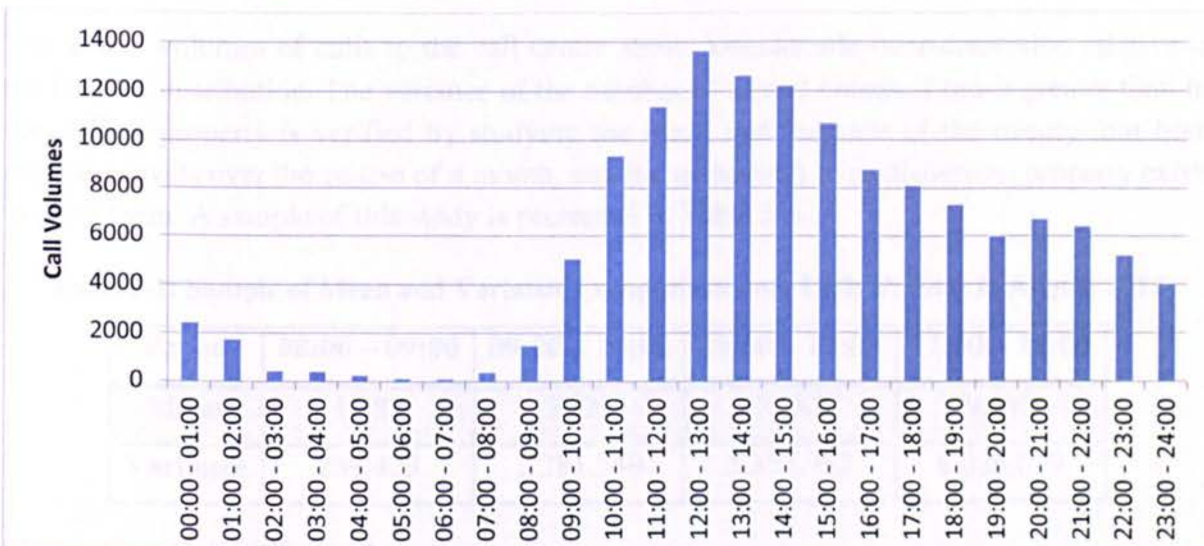


Figure 5-6: Frequency of hourly call arrivals of a typical day

In addition to period-to-period variability during the same day, call arrivals in a certain period experiences a significant variation from day to day. The following chart, Figure 5-7, shows the amount of variability over the twenty-four hour periods measured along one month period (August 2012). The inner line corresponds to the minimum volume in each period during that month. On the other hand, the outer line corresponds to the maximum volume in each period during the same month. Thus, the included region represents the significant amount of variation in each period from day to day over that month.

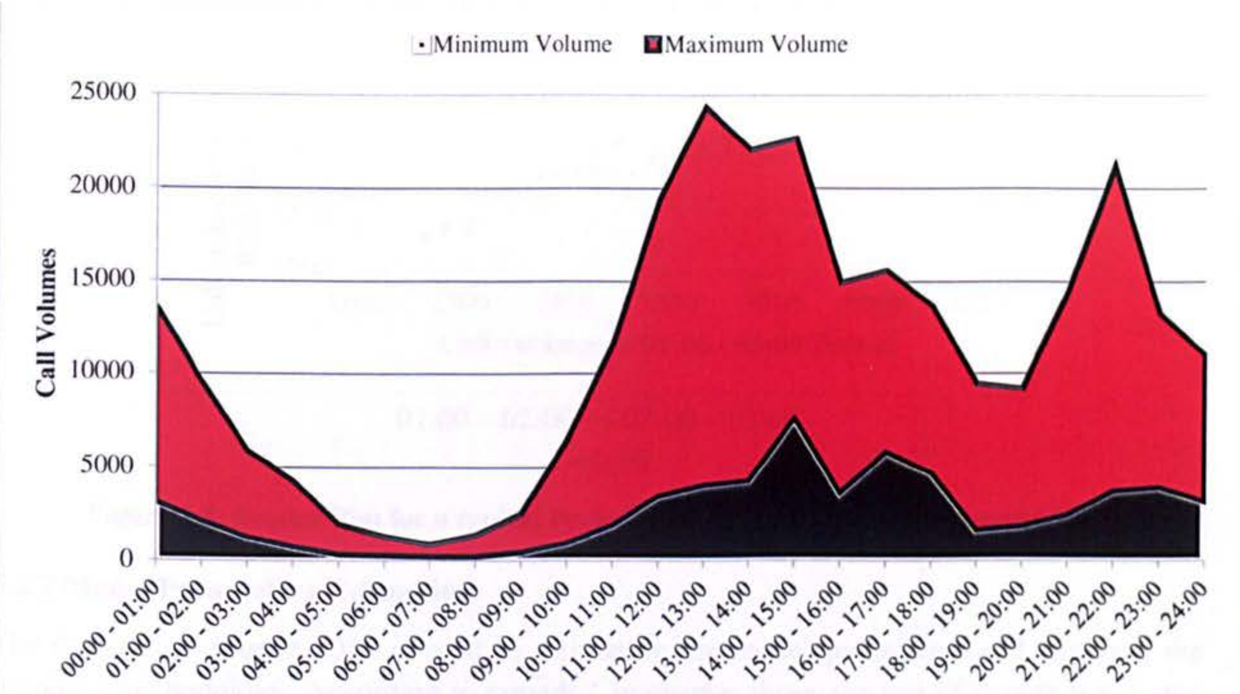


Figure 5-7: Variation in hourly period volume from a day to another in August 2012

Over-dispersion Property

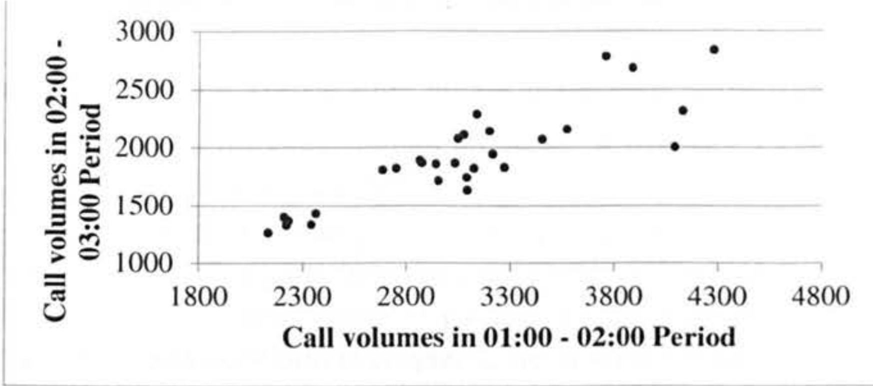
The arrival volumes of calls to the call centre show considerable over-dispersion relative to the Poisson distribution. The variance of the number of arrival counts is much greater than its mean. This property is verified by studying the mean and variance of the twenty-four hour periods arrivals over the course of a month, and the mentioned over-dispersion property exists in all of them. A sample of this study is presented in Table 5-1.

Table 5-1: Sample of Mean and Variance Comparison over Daily Periods in August 2012

Period	**08:00 – 09:00**	**09:00 – 10:00**	**10:00 – 11:00**	**11:00 – 12:00**
Mean	1,387	3,929	7,283	9,840
Variance	259,423	2,281,549	5,353,992	8,026,619

Strong Correlation Property

The detailed analysis of call arrivals data shows that there is an evident dependence between the arrival volumes of calls in the subsequent time periods during the day. One reason for this correlation may be due to the retrial phenomenon that occurs when callers – that abandoned in a period – retry to call again but in the subsequent period. This property is verified by calculating the *Pearson Correlation Coefficient* ***r*** for the successive daily-period pairs over the course of a month. The calculated values show a strong positive correlation between the subsequent periods where 83% of the twenty-four pairs possess a correlation coefficient greater than (0.5). Figure5-8 provides a typical scatter plot for the relation between call volumes in the subsequent periods showing the mentioned positive correlation.



01:00 – 02:00 → 02:00 – 03:00
$r = 0.95$

Figure 5-8: Scatter Plot for a typical Period-Pair Arrivals Volume in August 2012

5.2.2 Model Parameters Estimation

The data set of August 2012 is used in estimating the model parameters and applying the proposed methodology. According to remark 1 in chapter three, the test of hypothesis on the significance of the statistical difference between total daily volumes showed that the week days could be divided into four clusters in each of which the parameters of the normally distributed total daily arrivals are identical. This could be concluded from the statistical comparisons, between the total daily arrivals of different week days' pairs, made using two samples t-test. Knowing that the pair failing to reject the null hypothesis yields a high p-value greater than the significance level $\alpha = 0.05$, then we could deduce that Sunday, Monday, and Tuesday form a separate cluster as apparent from the p-values for the different pairs of these days as shown in Table 5-2. Additionally, Wednesday and Thursday give another cluster with a p-value of 0.56; while each of Friday and Saturday form separate clusters as they show very low p-values for all pairs with other weekdays. Likewise, referring to remark 2, all week days except Friday have the same structure of the daily period proportions. Friday shows a different structure of the proportions because Friday is the weekend vacation.

Table 5-2: p-values of t-test of total daily arrivals volumes pairs

	Saturday	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday
Saturday		0.001	0.001	0.001	0.053	0.261	0.000
Sunday	0.001		0.603	0.451	0.155	0.076	0.000
Monday	0.001	0.603		0.705	0.117	0.058	0.000
Tuesday	0.001	0.451	0.705		0.11	0.053	0.000
Wednesday	0.053	0.155	0.117	0.11		0.556	0.000
Thursday	0.261	0.076	0.058	0.053	0.556		0.000
Friday	0.000	0.000	0.000	0.000	0.000	0.000	

Gamma Dependent Random Arrival Rate Model Parameters Estimation

Following the parameter estimation algorithm discussed in section 2.3.1, the values of γ for each cluster could be estimated separately from equation 2-7, and λ could also be estimated for each period of the twenty-four daily periods in the four clusters from equation 2-8.This results in estimating twenty-five parameters for each cluster and summing up to a total of one hundred parameters. The estimated values of parameter λ_i for the first cluster, estimated by the presented and discussed algorithm in chapter 2, are provided in Table 5-3, while the value of γ equals 168.573014. In this table and the subsequent tables and figures, the twenty-four daily periods are ranked from 1 to 24 where 1 indicates period from 00:00 to 01:00 and 24 indicates period from 23:00 till 24:00.

Table 5-3: Estimated parameters of λ_i for the first cluster

Period (i)	λ_i	Period (i)	λ_i	Period (i)	λ_i
1	26.8078	**9**	8.9378	**17**	44.4713
2	18.4988	**10**	27.0387	**18**	35.1938
3	11.2446	**11**	49.4836	**19**	23.9777
4	5.8554	**12**	66.0746	**20**	31.4760
5	3.3430	**13**	77.1970	**21**	38.1829
6	2.3337	**14**	78.5317	**22**	38.4075
7	2.3966	**15**	64.2230	**23**	33.9295
8	3.5000	**16**	52.4928	**24**	31.3368

Daily Proportion-Based Arrival Rate Model Estimation

In this model, there are actually two models whose parameters need to be estimated, those of the model of total daily arrivals and those of the model of period proportion. In order to estimate the total daily arrivals model parameters, the mean and standard deviation for each cluster are calculated using equations 3-9 and 3-10 resulting in eight parameters (e.g., for cluster 1, μ = 127,249 and σ = 12,663). The parameters of the period proportion model are estimated according to the algorithm discussed in Section 3.2.1 (the estimation of the parameter α_i is performed using **R** software package).

Table 5-4: Estimated parameters α_i and β_i for weekdays Saturday – Thursday

Period (i)	α_i	β_i	Period (i)	α_i	β_i	Period (i)	α_i	β_i
1	30.6684	847.0611	**9**	10.3347	867.3948	**17**	50.6733	827.0562
2	21.3865	856.343	**10**	30.3106	847.4189	**18**	39.8127	837.9168
3	13.1201	864.6094	**11**	56.0716	821.658	**19**	24.5443	853.1852
4	7.1778	870.5517	**12**	74.8014	802.9281	**20**	35.2909	842.4386
5	4.2289	873.5006	**13**	87.5192	790.2103	**21**	43.1276	834.6019
6	3.1484	874.5811	**14**	89.5379	788.1916	**22**	42.4977	835.2318
7	3.1484	874.5811	**15**	72.7849	804.9446	**23**	38.4475	839.2821
8	4.3651	873.3644	**16**	59.1776	818.5519	**24**	35.5541	842.1754

5.2.3 Models Validation

The main properties of the studied/developed models are investigated and compared to the real-data main properties to quantify how far they are valid and represent the real data. This is done by sampling from the developed models and comparing how close the generated samples are to the real data set. The following results are obtained from the analysis of the first cluster data. The first thing to be compared is the estimated and real means of arrival counts during the different periods. The average arrivals comparison within different periods of a typical month shown in Figure 5-9 shows that the developed models succeed in producing similar estimates for the arrival counts. Also, the difference between both models (the Gamma-dependent model and the proposed one) is negligible.

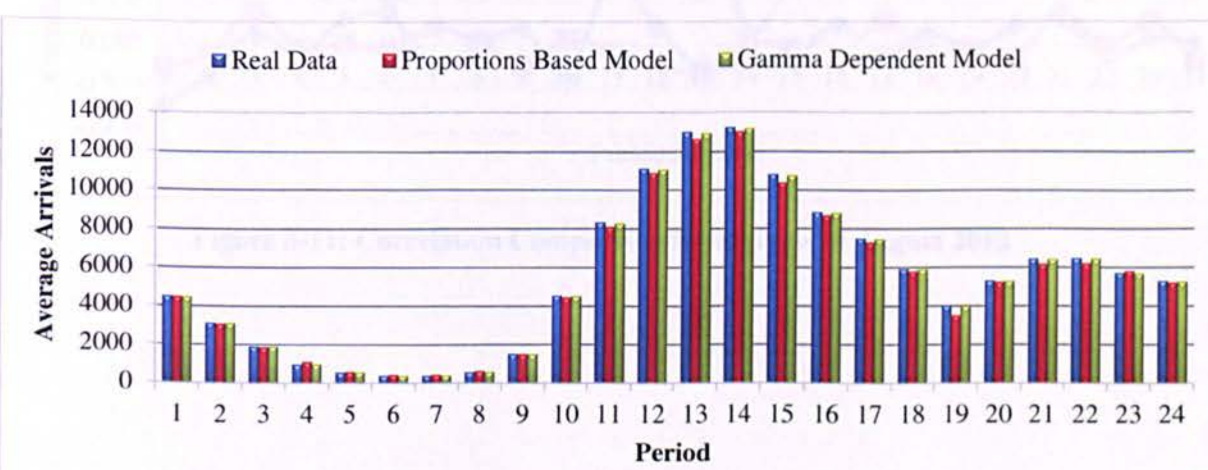


Figure 5-9: Comparison of Estimated and Real Means of Call Volumes/Period of August 2012

Comparing the coefficient of variation (CV), as shown in Figure 5-10, reveals that the Gamma dependent model fails in showing a coefficient of variation close to the real one. It, nearly, shows a zero value for the CV. On the other hand, the proportions-based model outperforms the other model in the CV comparison. The estimated CV values match that of real data in the periods between period starting at 09:00 and ending at 18:00 and from 00:00 to 04:00. However, the estimated CV fails to do that in the remaining periods.

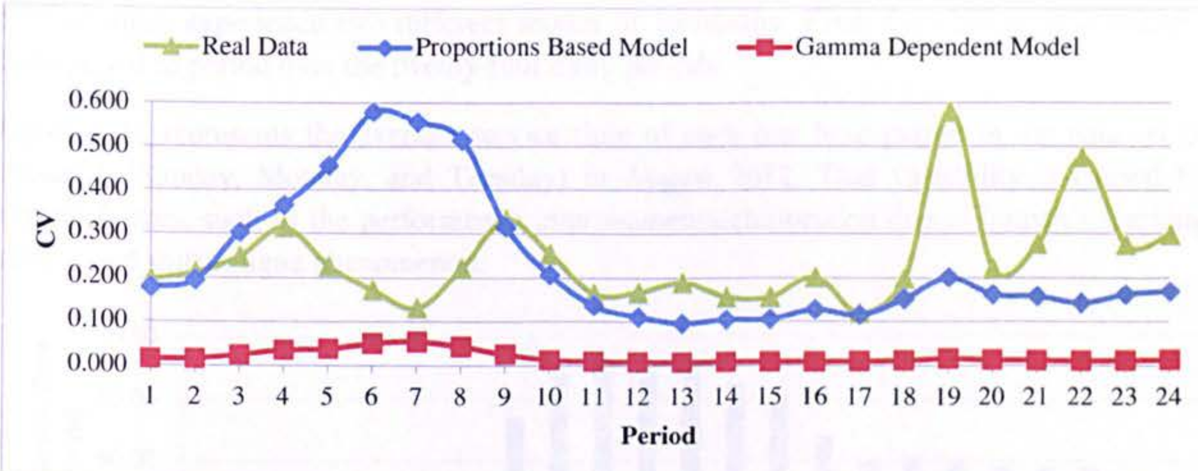


Figure 5-10: Coefficient of Variation Comparison for arrivals of August 2012

As shown figure 5-11, the developed models produce nearly similar correlation. However, these structures are far from the real one and produce several negative correlation values.

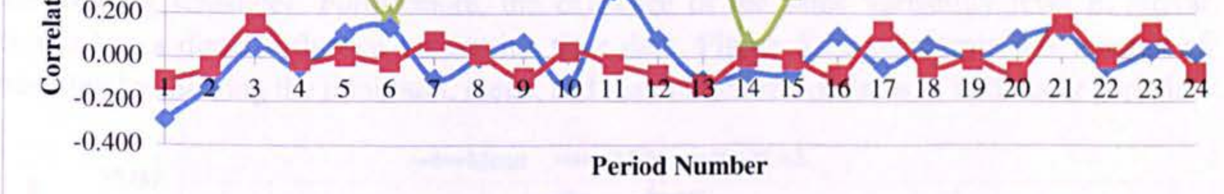


Figure 5-11: Correlation Comparison for arrivals of August 2012

5.3 Service Time Model Implementation

The available empirical data about service times, from the call centre, was the average handling time for the arriving calls in each one hour-period. In this study, the same four-clusters of data used in modeling the arrival counts is used here to work on developing the service time models.

5.3.1 Prelude Data Analysis

Service times experience two different modes of variability. First, the service time changes from period to period over the twenty-four daily periods.

Figure 5-12 represents the average service time of each one hour-period in the data set of cluster 1 (Sunday, Monday, and Tuesday) in August 2012. That variability is caused by several factors, such as the performance improvement/deterioration due to human's learning ability, and shift-fatigue phenomenon.

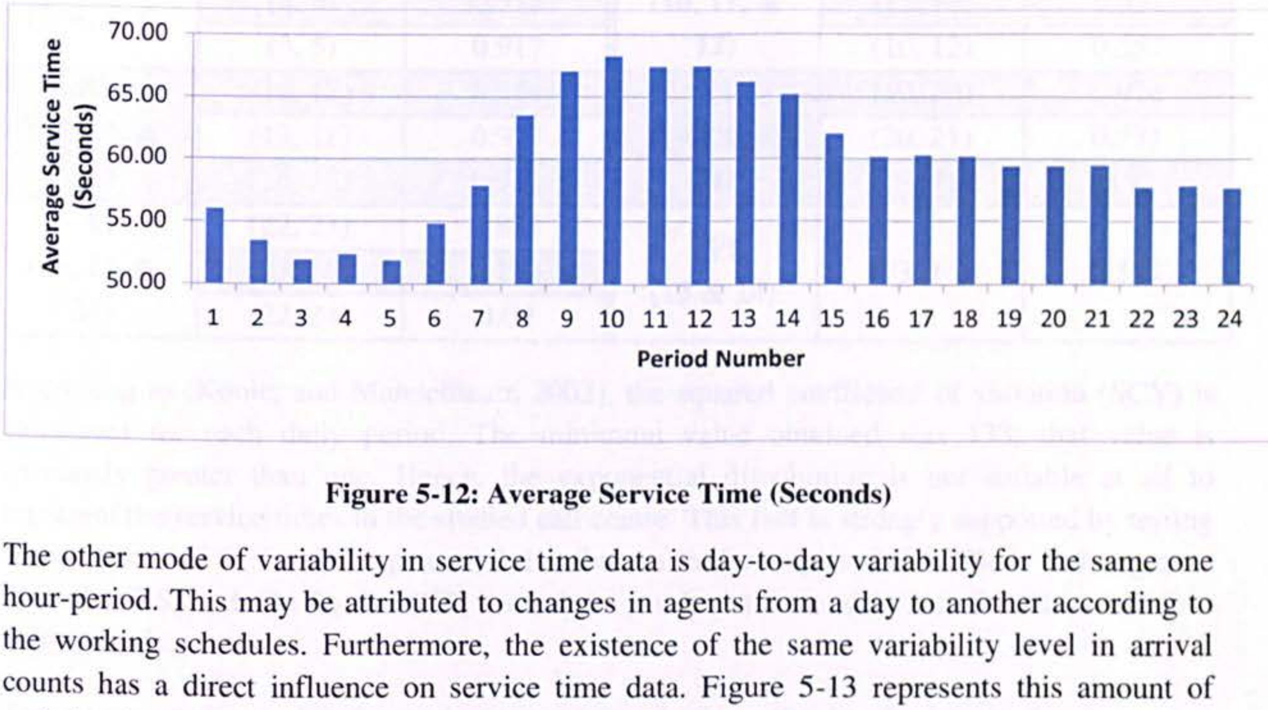


Figure 5-12: Average Service Time (Seconds)

The other mode of variability in service time data is day-to-day variability for the same one hour-period. This may be attributed to changes in agents from a day to another according to the working schedules. Furthermore, the existence of the same variability level in arrival counts has a direct influence on service time data. Figure 5-13 represents this amount of variation by showing the minimum, mean, and maximum service times of each daily period.

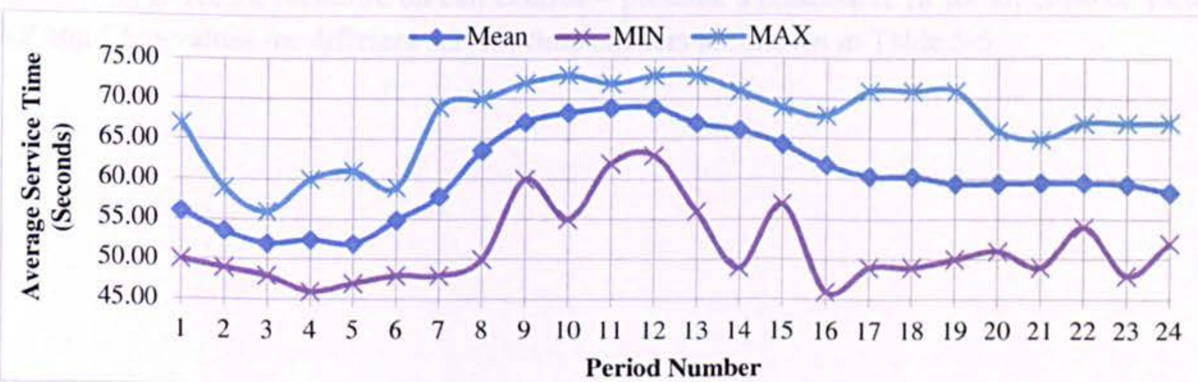


Figure 5-13: Profile of Minimum, mean, and maximum service times

5.3.2 Model Parameters Estimation

In Figure 5-12, one could assume that there are some service time clusters of the different periods. Periods 22, 23, and 24, for example, show nearly the same average over the available days in the data set. This assumption could be supported by testing the significance of statistical difference between mean service time data sets of the different daily periods. These statistical comparisons – done by two samples t-test – reveal five clusters each of which contains three-periods, one cluster of two-period size, and the others are single-period clusters. The single period clusters contain, separately, periods 1, 2, 6, 7, 8, 9, and 15. The other clusters are presented in Table 5-5.

Table 5-5: Statistical comparisons between service time data sets

Cluster	Pairs Comparison		Cluster	Pairs Comparison	
	Pair	p-Value		Pair	p-Value
I. (3, 4, & 5)	(3, 4)	0.516	**II. (10, 11, & 12)**	(10, 11)	0.613
	(4, 5)	0.517		(11, 12)	0.951
	(3, 5)	0.917		(10, 12)	0.557
III. (16, 17, & 18)	(16, 17)	0.311	**IV. (19, 20, & 21)**	(19, 20)	0.974
	(17, 18)	0.977		(20, 21)	0.931
	(16, 18)	0.312		(19, 21)	0.92
V. (22, 23, & 24)	(22, 23)	0.873	**VI. (13 & 14)**	(13, 14)	0.394
	(23, 24)	0.134			
	(22, 24)	0.09			

According to (Koole, and Mandelbaum 2002), the squared coefficient of variation (SCV) is calculated for each daily period. The minimum value obtained was 138; that value is extremely greater than one. Hence, the exponential distribution is not suitable at all to represent the service times in the studied call centre. This fact is strongly supported by testing the goodness of fit for the exponential distribution to the empirical data. Both Kolmogorov-Smirnov (KS) and Chi-Square (CS) tests show p-values that tend to zero for all twenty-four daily periods.

Testing the goodness of fit for the log-normal distribution – the distribution that has been used extensively in recent literature on call centres – presents a reasonable fit for all clusters. Both KS and CS p-values for different service time clusters are shown in Table 5-6.

Table 5-6: KS and CS P-values for fitting the log-normal distribution to service times

Cluster Periods	KS P-Value	CS P-Value	Cluster Periods	KS P-Value	CS P-Value
1	0.34	0.38	**10, 11, 12**	0.02	0.8
2	0.21	0.32	**13, 14**	0.001	0.18
3, 4, 5	0.15	0.25	**15**	0.52	0.67
6	0.2	0.76	**16, 17, 18**	0.04	0.08
7	0.46	0.92	**19, 20, 21**	0.02	0.33
8	0.36	0.09	**22, 23, 24**	0.43	0.04
9	0.51	0.12			

It is found that it would be more accurate to use a shifted log-normal distribution. The estimated values of log-normal distribution parameters and the shifting constant are shown in Table 5-7. These parameters are estimated using *Arena® Input Analyzer* software package.

Table 5-7: Parameters of log-normal distribution for service time

Period (i)	Constant	Log μ	Log σ	Period (i)	Constant	Log μ	Log σ
1	43	9.27	3.90	**13, 14**	51	15.70	6.58
2	46	7.76	3.22	**15**	56	8.55	3.68
3, 4, 5	43	9.28	3.93	**16, 17, 18**	45	16.30	7.26
6	45	10.10	3.61	**19, 20**	48	11.60	4.69
7	48	10.30	5.13	**21**	50	11.6	4.69
8	47	16.80	5.17	**22**	49	10.20	4.51
9	58	9.24	3.61	**23, 24**	48	10.20	4.51
10, 11, 12	60	9.00	2.58				

5.4 Abandonment Volumes Model Implementation

In order to model the patience of customers and their abandonment behavior, the form of historical data available regarding this point is an instrumental input to the model construction. Usually, this behavior is modeled by assigning a patience time attribute to each customer. Afterwards, if that patience time is exceeded and the customer is still in the queue, it will abandon the queue directly. The form of available empirical data, however, did not contain such records of patience time of individual calls or even aggregated averages. Alternatively, the available data is the volumes of abandoned customers within separate six patience-time slots. Therefore, the abandonment behavior is modeled using a couple of parameters, as discussed in section 3.2.2, the patience time and the abandonment group weight/percentage.

5.4.1 Patience Time

Based on the available data, there are six separate patience time slots: zero-second (customers abandon the queue when they know that they will wait), 1 to 10 seconds, 10 to 20 seconds, 20 to 30 seconds, 30 to 60 seconds, and *maximum-patience-time* slot. The patience time for the first group is considered to be zero second (i.e., the customer abandon when just arrive the system). The patience time for customers classified to be in the second, third, fourth, and fifth groups is modeled as a uniform random variable. The parameters of the uniform distribution for each slot is the start and end time values of the time slot itself. The patience time of group six (*maximum-patience-time*) is modeled using a triangular distribution. Table 5-8 shows the parameters of that triangular model for each time period.

Table 5-8: Triangular distribution parameters of *maximum-patience-time*

Period (*i*)	Min	Mode	Max	Period (*i*)	Min	Mode	Max	Period (*i*)	Min	Mode	Max
1	64	74	90	**9**	11	17	62	**17**	12	59	106
2	62	77	80	**10**	12	13	24	**18**	11	17	60
3	40	71	81	**11**	12	37	96	**19**	12	17	63
4	11	16	66	**12**	50	62	109	**20**	10	11	13
5	11	12	23	**13**	56	91	173	**21**	10	14	60
6	11	15	35	**14**	29	89	230	**22**	10	14	60
7	11	30	64	**15**	11	47	74	**23**	12	22	93
8	12	60	66	**16**	25	67	94	**24**	11	72	79

5.4.2 Abandonment Group Proportion

According to the proposed model, each one-hour period has six abandonment groups divided based on their proportions in the total number of arrivals in that hour. Therefore, a six-variable Dirichlet random vector is required for each one-hour period to represent these proportions. The parameters of this random vector are estimated using **R** statistical package. The estimated values for each hour are presented in Appendix C.

5.4.3 Model Validation

It is found that the developed abandonment behavior model provides a realistic representation of the abandonment volumes. This is proven by applying the validation methodology, discussed in section 4.2, to compare the real data of total abandonment volumes in each hour and the simulated data of the same parameter. According to the procedure of that methodology, all of simulated data sets cannot be considered normally distributed except the data set of period 10. Therefore, the non-parametric mean equality test, *Wilcoxon-Mann-Whitney*, is used to compare the real and simulated data sets.

The results, shown in Table 5-9, of that test prove the validity of the proposed model through showing high p-values for most of the twenty-four periods' abandonment data.

Table 5-9: Validation of abandonment volumes model

Statistical Test	Normality Test *Anderson-Darling-test*		Mean Equality *Wilcoxon-Mann-Whitney*
Period No.	**Real Data p-value**	**Simulated Data p-value**	**p-value**
1	0.963	0.005	0.4350
2	0.385	0.005	0.6145
3	0.008	0.005	0.0742
4	0.182	0.005	0.9735
5	0.474	0.005	0.2367
6	0.034	0.005	0.2416
7	0.266	0.005	0.0910
8	0.005	0.005	0.0160
9	0.748	0.030	0.7872
10	0.283	0.210	0.2334
11	0.087	0.005	0.0166
12	0.058	0.005	0.0535
13	0.139	0.005	0.4880
14	0.064	0.005	0.8779
15	0.005	0.005	0.9206
16	0.005	0.005	0.1350
17	0.114	0.005	0.0226
18	0.005	0.005	0.7366
19	0.005	0.011	0.0325
20	0.321	0.005	0.2383
21	0.005	0.005	0.8194
22	0.005	0.005	0.5554
23	0.005	0.005	0.7523
24	0.099	0.005	0.8976

5.5 Conceptual Model Construction

The objective of this part of the work is to develop a simulation model that mimics the call centre stochastic operations. The developed simulation model is to be used as an evaluation tool for the service system performance of the studied call centre. The performance measures that could be evaluated include a set of output parameters, as described in section 2.1.4, such as *average speed of answer* (ASA), *service level* (SL), *abandonment percentage* (ABAND%), and *agent occupancy.*

Moreover, this tool is utilized to apply some sort of what-if-analysis to allocate the optimal staffing levels such that some service requirement constraints are satisfied. The selection of the constraints is based on the adopted management regime whether it is *quality driven regime, efficiency driven regime,* or *quality efficiency driven* (QED).

5.5.1 Model Logic

The call centre system behavior, discussed in section 2.1.3, is represented in detail in a flow chart shown in Figure 5-14. The model starts and proceeds, according to an entity-based logic, in a sequence of events characterized by a set of attributes and variables assigned to the entity (customer). Those events and their corresponding attributes are described as follows:

1. ***Arrival Event.*** The model starts by generating customers' arrivals from an infinite calling population. The arrival time of a specific entity is scheduled based on an attribute (drawn from a probability model) of the *inter-arrival time* between successive entities.
2. ***Welcoming Event.*** All arriving customers listen to a welcome message before they are routed to their service agent or to their position in the tele-queue. The time of the *welcoming duration* action is determined by a constant variable (since it is a fixed message) for all customers.
3. ***Routing Event.*** The *automatic call distributer* ACD searches the agents' pool for an idle ready agent to receive an incoming call. This is done through checking the value of a *binary state variable* assigned to each agent (i.e., one means busy; zero means idle). If the ACD finds more than one idle agent, it routes the call to the agent having the longest idle time (another agent's state variable). Otherwise, the ACD routes the customer to a tele-queue to wait for any agent to be idle and meanwhile he/she may listen to a commercial. The ACD gives each agent a four-second break after each call considering him/her as a busy agent during this break (i.e., the ACD does not route any calls to the agent or consider him/her as an idle agent unless this short break ends). This break is represented by a variable called *short break time.*
4. ***Waiting Event.*** The customers will be waiting in the tele-queue until one of the agents becomes idle. The position of the customer in the queue is a *decreasing variable* assigned to each customer during his/her waiting in the queue. Another attribute is assigned to each entity, at the beginning of this event, to represent the time interval between the start and the end of this action (i.e., measure for the *queue waiting time*).

5. *Service Event.* If there are any idle agents, then the customer will acquire the service directly. Otherwise, if the customer is patient enough until one of the agents becomes idle, he/she will be routed directly to the first idle agent. Each customer at this instant would be handled in a specific handling time. This *handling time* is an attribute assigned to each entity and drawn randomly from a probability distribution characterizing the service activity. Moreover, this event is also characterized by the capacity variable that determines the number of the available agents to provide the acquired services. In the test runs and model validation stage, the *capacity variable* is drawn randomly from a probability distribution. In the phase of production runs and analysis, this would be the decision variable in the optimization context.

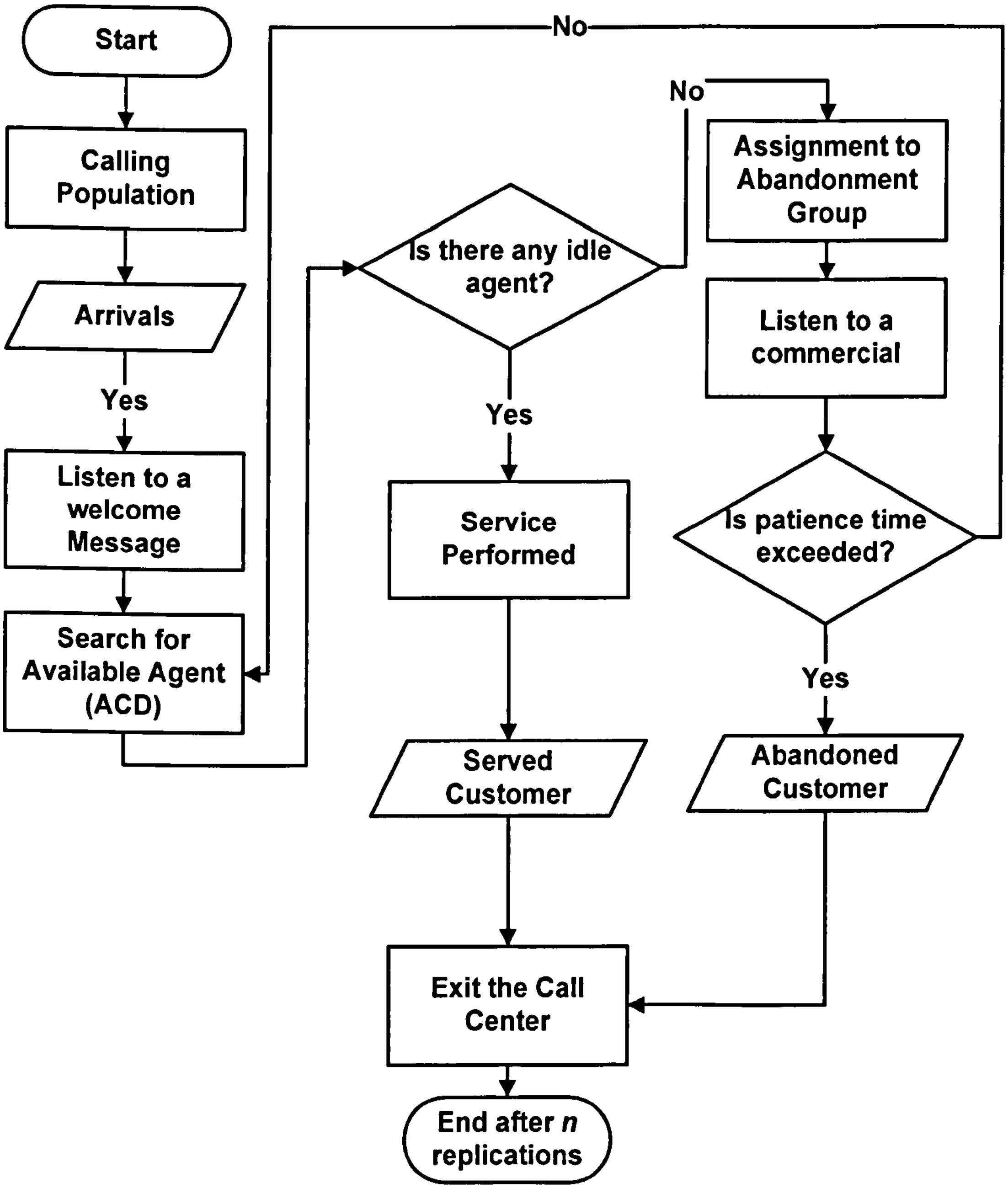


Figure 5-14: The model logic of the call centre dynamics

6. ***Abandonment Event.*** It is an alternative event to the event number five where the customer is not patient enough to wait for the service. The impatient customer would abandon the queue at any instant and be considered as a lost call. The details of this event and its logic will be described separately in the next section 5.5.2.

7. ***Departure Event.*** All customers arriving to this call centre experience this event. They leave the system in one of two possible states either served or abandoning.

5.5.2 Patience and Abandonment Logic

All arriving customers are classified into six-abandonment groups according to their patience-time slot. This classification is based on the weight of each group relative to the total number of arrivals. Thus, two attributes are assigned to each customer to model this logic. The first attribute, *abandonment classification attribute*, is an integer value which varies between 1 and 6 assigned based on the weight of each abandonment group. The second attribute is the *patience time attribute* that each customer would experience. The patience time is a random variable that varies between the bounds of each time slot in the case of the first five groups. As for the sixth group, the patience time is represented as a random variable modeled from the maximum patience times recorded in the system.

5.6 Random Numbers Generation

In this step, the method to generate random numbers, according to the aforementioned data models, representing the attributes and variables chosen in the previous step is discussed. The focus of the study presented here is on cluster 1 that consists of Sunday, Monday, and Tuesday so that all used values correspond to this cluster.

1. ***Inter-arrival Time.*** The inter-arrival time between the successive calls is assumed to be exponentially distributed. The rate of the exponential distribution is a stochastic random variable that changes continually with the time through the day. Thus, this rate is considered as an hourly changing percentage of the total volume of daily arrivals. This percentage is represented by a Dirichlet distribution and generated for each hour via a suitable Beta distribution (as described in section 4.2). Moreover, the total daily volume of arrivals is normally distributed. Then, the used function to generate the random inter-arrival time attribute $T_{arr}(i)$, in minutes, is

$$T_{arr}(i) = EXPO\left(60/\left(BETA(\alpha_i, \beta_i) * NORM(\mu, \sigma)\right)\right) \quad 5\text{-}1$$

2. ***Welcoming Duration.*** This duration (T_{wel}) is considered to be constant for all arriving entities and its value is *thirteen* seconds.

3. ***Short Break Time.*** This duration (T_{break}) is considered to be constant for all arriving entities and its value is *four* seconds.

4. ***Handling Time.*** The handling time attribute $T_{serv}(i)$ assigned for each customer is log-normally distributed shifted by a constant value. This attribute also exhibits a stochastic

nature as it changes continually on an hourly basis. The function that is used in random number generation for this attribute is given as follows

$$T_{serv}(i) = Constant\ (i) + LOGNORM(\log \mu_i, \log \sigma_i) \qquad 5\text{-}2$$

5. ***Service Capacity.*** In order to validate the model, the number of scheduled agents *N* (*i*) should be considered as an input variable to the model. Thus, it is modeled as a random variable for each daily period. This random variable follows either a uniform distribution or a triangular distribution determined according to the real scheduled agents' levels in the different days of the available empirical data set. The parameters of the service capacity distribution of each period are presented in the following table (a period that does not have a mode of its capacity data is modeled via a uniform distribution).

Table 5-10: Parameters of service capacity variable

Period (*i*)	Min	Mode	Max	Period (*i*)	Min	Mode	Max	Period (*i*)	Min	Mode	Max
1	43	50	80	**9**	100	---	180	**17**	130	140	200
2	38	45	65	**10**	130	150	165	**18**	125	160	165
3	23	---	47	**11**	143	---	200	**19**	95	---	185
4	22	---	50	**12**	160	217	235	**20**	100	110	140
5	18	27	38	**13**	180	230	248	**21**	85	110	130
6	16	28	31	**14**	190	244	263	**22**	85	110	130
7	11	14	24	**15**	160	230	245	**23**	74	90	119
8	15	18	44	**16**	80	190	250	**24**	56	63	116

6. ***Abandonment Classification.*** Arriving customers are classified into six groups based on their abandonment category through an integer variable, *abandonment group index b*, varying between 1 and 6. Each group contains a percentage R_b (*i*) of the total number of arrivals; these percentages for each daily period are stochastic variables that are represented by a Dirichlet distribution and generated for each period via a Beta distribution (as described in section 3.2.1).

7. ***Patience Time.*** The patience time attribute, $T_{patience}$ (*b*), is modeled by a uniform distribution for the second, third, fourth, and fifth patience-time-slots. The parameters of this distribution (minimum and maximum values) are the bounds of time of slot of the four slots. The patience time for the *maximum-patience-time-slot* is modeled using triangular distribution.

5.7 Model Translation into a Computerised Model

Computer simulation refers to methods for studying the models of real world systems by numerical evaluation using computer processed logic. In order to translate the model logic of the studied call centre into a computerised model, the software package *Arena*® provided by Rockwell Automation is used.

5.7.1 Modelling Approach

The model starts by creating the customers' arrivals to the call centre. This is done based on the exponential inter-arrival time attribute T_{arr} (i) whose rate changes stochastically every period i through the day. Those arrivals enter a service system of a single *first-in-first-out* queue with multiple servers of N-uniformly distributed size. The service provided by those N (i) servers requires a random handling time T_{serv} (i) that is log-normally distributed. Likewise, both service capacity N (i) and handling times T_{serv} (i) change stochastically every period i through the day.

In order to represent abandonments, the impatient customers are willing to wait only a limited time period before they abandon the queue. The patience time tolerance, $T_{patience}$ (b), will be generated randomly from a probability distribution that changes stochastically based on the abandonment group (b) that each customer is assigned to.

Modeling the dynamics of the abandonment activity represents a real challenge. If the arriving call is allowed to enter the queue and the patience time is reached, then that call would need to be found and removed from the queue. Thus, from the modeling perspective, the arriving customer must be placed in the waiting line and the abandonment event should be performed after the patience time has elapsed. If the customer is, however, actually placed in the queue, there is no direct built-in mechanism to detect that the patience time has elapsed. To overcome this obstacle, a duplicate of each arriving entity, that carries all original entity's characteristics (attributes), is created, but follows a different logic than the one followed by the original entity. The original entity, which represents the actual customer, is sent to the service queue. The duplicate entity, on the other hand, will be delayed for an amount of time that equals its patience time. After the patience time delay, the duplicate entity searches the waiting line for its original entity. If the customer is no longer in the service queue (i.e., the customer has been already served), then the duplicate entity will be disposed from the system. Otherwise, the customer is still in the queue, the duplicate entity will remove the original entity from the queue and dispose of both itself and the original entity. This model logic is outlined in the form of a pseudo-code as shown in Figure 5-15. In order to implement this modeling approach, the operational data structures should be defined first before building the entity-based logic. The following section focuses on finding a suitable representation for the used data sets while section 0 is concerned with the logic building part.

5.7.2 Stochastic Data Structure

Due to the stochastic nature of the call centre, most of used attributes change on an hourly-basis through the working day. This is represented by building schedules for each one of those attributes. Those schedules contain twenty-four values for each attribute where each value corresponds to a daily-one-hour period. Moreover, the schedule is repeated at the beginning of a new simulated day. The attributes that are modeled using the schedules are represented as shown in Table 5-11.

Table 5-11: Used schedules and the corresponding attributes

Schedule Name	Corresponding Attribute
Proportions Part in Inter-arrivals Time	$P_i = BETA(\alpha_i, \beta_i)$
Abandonment Percentage Zero-Second	$R_1(i)$
Abandonment Percentage 1 to 10 Sec.	$R_2(i)$
Abandonment Percentage 10 to 20 Sec.	$R_3(i)$
Abandonment Percentage 20 to 30 Sec.	$R_4(i)$
Abandonment Percentage 30 to 60 Sec.	$R_5(i)$
Abandonment Percentage Greater than 60 Sec.	$R_6(i)$
Patience Time for maximum-patience-time-group	$T_{max}(i)$
Handling Time	$T_{serv}(i)$
Agents Schedule (Capacity Variable)	$N(i)$

Create Arrivals with $T_{arr}(i)$ = EXPO ($60/\lambda_i$); λ_i = BETA (α_i, β_i) * NORM (μ, σ)
    Record Number of Arrivals: Number of Arrivals = Number of Arrivals + 1
    Assign Abandonment Group Index b = DISC (cumulative $R_b(i), b$); for b = 1 to 6 and i = 1, 2... 24)
        b = DISC [$R_1(i)$, 1; $\sum_{b=1}^{2} R_b(i)$, 2; $\sum_{b=1}^{3} R_b(i)$, 3; $\sum_{b=1}^{4} R_b(i)$, 4; $\sum_{b=1}^{5} R_b(i)$, 5; 1.0, 6]
            $R_b(i)$ = BETA ($\alpha_b(i), \beta_b(i)$)
            $R_6(i) = 1 - \sum_{b=1}^{5} R_b(i)$
    Assign Patience Time
        If $1 < b \leq 5$, then $T_{patience}(b)$ = UNIF ($t_s(b), t_e(b)$)
        ELSE, b = 6, then $T_{max}(i)$ = TRIA (MIN $T_{max}(i)$, MODE $T_{max}(i)$, MAX $T_{max}(i)$)
    Assign Entity Arrival Time: Enter System = Time in Simulation Clock Now (TNOW)
    Assign Entity Serial Number: Customer # = Customer # + 1
    Decide Whether Zero-Patience Entity
        If Abandonment Group Index b = 1
            Record Number of Abandoned Customers: Abandoned Customers + 1
            Dispose Zero-Patience Entity
        ELSE, Delay by Welcoming Duration T_{wel} = 13 seconds

Create Duplicate Entity 1
    Route Original Entity to Servers Queue
        Assign Queue Entry Time: Queue Entry Time = TNOW
        Wait till any agent's state variable = 0
        Tally Queue Waiting Time: Queue Waiting Time = TNOW – Queue Entry Time
        Route Original Entity to Idle Server & Let agent's state variable = 1
        Delay for Handling Time $T_{serv}(i)$ = Constant (i) + LOGNORM (μ_i, σ_i)
    Create Duplicate Entity 2
        Duplicate Entity 2
            Delay for Break Time T_{break} = 4 seconds
            Let agent's state variable = 0
        Original Entity
            Tally System Time: System Time = TNOW – Enter System
            Record Number of Served Customers = Number of Served Customers + 1
            Dispose Original Entity
    Route Duplicate Entity to Abandonments Control Logic
        Delay for Patience Time $T_{patience}(b)$
        Assign Search Attribute: Search # = Customer #
        Search Servers Queue for Original Entity
            If Search # = Customer # (of Waiting Original Entity)
                Assign Rank of Entity Found J to Duplicate Entity
            ELSE, Dispose Duplicate Entity
        Remove Original Entity of Rank J from Servers Queue & Dispose It
        Record Number of Abandoned Customers: Abandoned Customers + 1
        Dispose Duplicate Entity

Figure 5-15: Pseudo-code of the simulation model logic

5.7.3 Computational Logic

The main logic of the computational model of the system consists of the main events occurring in the real system. Those events are customers' arrivals, welcoming event, waiting event, service event, abandonment event, and departure event. Each event of those main events has been treated as a sub-system so that each one of them has its own computational modeling logic in the proposed Arena model, as shown in Figure 5-16.

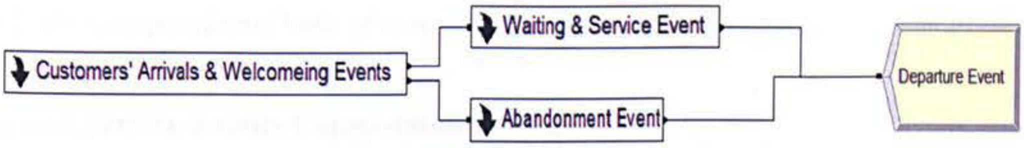


Figure 5-16: Computational model of system as presented in *Arena*®

Customers' Arrivals and Welcoming Events Computational Logic

The first sub-system logic, customers' arrivals and welcoming events, is shown in Figure 5-17. The model starts with a create model that creates arriving customers with inter-arrival time, in minutes, generated according to equation 5-1; $(T_{arr}(i) = EXPO\left(60/(BETA(\alpha_i, \beta_i) * NORM(\mu, \sigma))\right)$.

The created arrivals pass through the record module to increment the value of the number of arrivals variable by one. After that, the entities are sent directly to the assign module where the values of the main attributes, required later, are assigned to each entity. Those assignments are abandonment group index, patience time, entity arrival time, and entity serial number. At this instant, the zero-patience customers (who decide to leave the call centre when they just connected to the call centre) will be separated from the arrivals stream to abandon the call centre. This separation is modeled by using a decide module that identifies them by checking whether the abandonment group index equals one or not. The zero-patience customer will be sent to a record module to increment the variable number of abandoned customers by one. This is done using a route module that transfers the entity to that record module in the abandonment event logic. The second real event (welcoming event) will be modeled using a delay module to delay each entity by four seconds. The delayed entities are then sent to the separate module. This module is responsible for making the duplicates of the original entities. The original entity leaves the separate module to start the waiting event logic. On the other hand, the duplicate entity goes directly into the abandonment event logic.

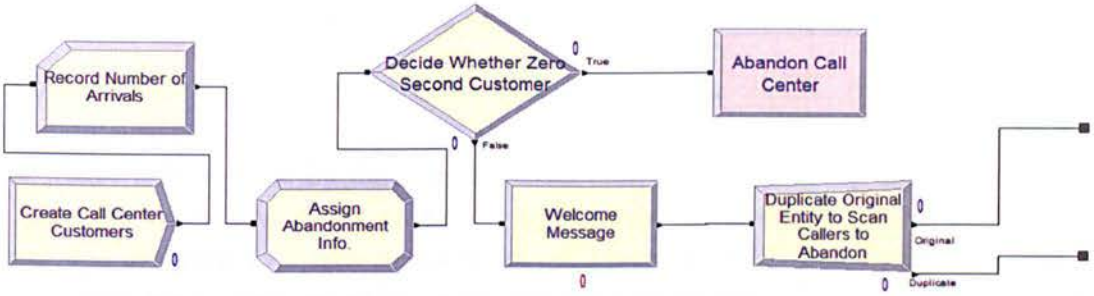


Figure 5-17: Computational logic of customers' arrivals and welcoming events as presented in *Arena*®

Waiting and Service Events Computational Logic

The original entity then proceeds through the logic of the waiting event, in the second sub-system shown in Figure 5-18, by recording queue entry time in an assign module. After this assignment, the entity is sent into a seize module which performs the task of the automatic call distributor ACD. This module checks the agent's state variable; if any agent has an instantaneous zero-value of this variable, then the entity will be sent to that agent passing first by a record module that evaluates the queue waiting time (zero in this case). If all agents are busy, then the entity will be sent into a queue module (embedded in the seize module) to model the waiting event. When one of the agents becomes idle, the entity will be sent to the record module to evaluate the waiting time. After that, the service event is modeled by a delay module where the entity will be delayed by the handling time. The agents are modeled by a resource module that has a stochastic capacity varying every hour based on a schedule named Agents Schedule. The served entity is then duplicated by a separate module. The original entity will be sent to two consecutive record modules; one to tally the system time and the second one to increment the value of the variable number of served customers by one and then will dispose the system. The second duplicate of the original entity is used to reserve the agent for an extra four seconds through using a delay module. Afterwards, the agent will be released by sending the duplicate entity to a release module.

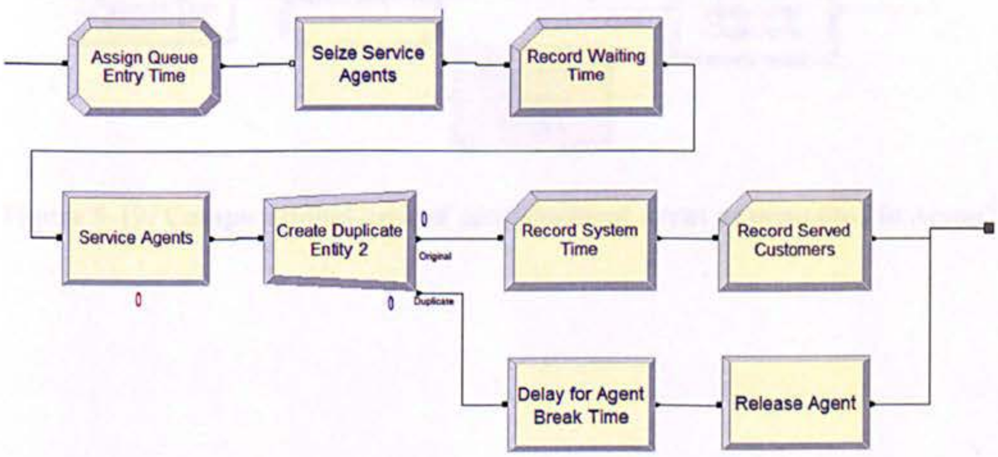


Figure 5-18: Computational logic of waiting and service events as presented in *Arena*®

Abandonment Event Computational Logic

The duplicate entity is sent to the third sub-system in the model, the sub-system for the abandonment logic. It starts by sending the first duplicate of the original entity to a delay module where it is delayed by the patience time that was assigned by the attribute patience time. After the delay, the entity enters an assign module where the value of the attribute customer # is assigned to a new variable named search #. After that assignment, the entity starts searching for its original entity in the servers waiting line. This is modeled by sending the entity to a search module that allows searching queues to find the rank, or queue position, of an entity that satisfies a defined search condition. The search module will search all entities in the servers queue for the entity that has the same value for its customer # attribute as the variable search # that is just assigned. After the search is performed, if the original entity, or customer, is no longer in the queue (i.e., the search condition is not satisfied), the duplicate entity will be disposed out of the system. If the original entity is found, its rank in the queue will be saved to a variable named J and the duplicate entity will be sent to a remove module. The remove module is used to model the abandonment action of the original customer by removing the entity of the rank J from the servers queue. At this point of the logic, the duplicate entity will be sent directly out of the system while the original entity will pass through a record module to increment the variable number of abandoned customers by one (this record module also receives the zero-patience entities through a station module). Finally, the original entity leaves the system through the dispose module which models the real call end.

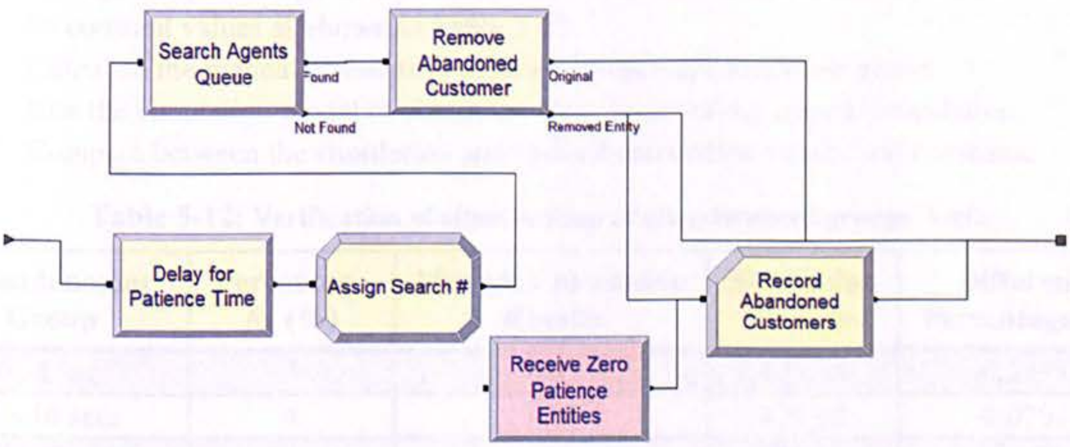


Figure 5-19: Computational logic of abandonment event as presented in *Arena*®

5.8 Model Verification

The logic of the computational model is verified through checking its ability to represent the main components of the proposed modeling approach (presented in section 5.7.1) which is derived from the conceptual model. In essence, the model is run using input parameters having some corresponding output values (could be computed by hand). Therefore, if the model results match already known results, then the developed model is verified. The main modeling issues to be checked are the classification of the arriving customers into the different abandonment groups, the abandonment action logic (search and remove), the processing sequence, and the generated random input parameters.

The issue of verifying that the random number generators produce input data streams that conform to the real data used in input modeling has been discussed in sections 5.2 5.3, and 5.4. The remaining modeling issues to be verified are discussed in the following sections.

5.8.1 Classification of Abandonment Groups

In order to test the proposed logic to classify arriving customers among the six abandonment groups, the following steps are applied to the proposed simulation model (the results are presented in Table 5-12:

1. Make the replication length correspond to a maximum number of arrivals of 12,000 customers.
2. Change the percentage of each abandonment group in the total number of arrivals to be constant values as shown in Table 5-12.
3. Calculate the manual calculation volumes of each abandonment group.
4. Run the simulation model to obtain the same result of the manual calculation.
5. Compare between the simulation and manual calculation results and comment.

Table 5-12: Verification of classification of abandonment groups' logic

Abandonment Group	Percentage R_b (%)	Manual Calculation Results	Simulation Results	Difference Percentage (%)
0 - 1 sec.	4	480	478.18	-0.37917
1 - 10 sec.	4	480	479.62	-0.07917
10 - 20 sec.	3	360	361.35	0.375
20 - 30 sec.	2	240	238.35	-0.6875
30 - 60 sec.	6	720	723.51	0.4875
Maximum-Patience-Time	81	9720	9718.99	-0.01039
Total	100	12000	12000	0

The simulation results are obtained by running the simulation model for 100 replications and taking their average. It should be noticed that those results are not typically the same as the by-hand-results. Consequently, the used mechanism in the simulation is to assign arriving customers to the different groups randomly and then modify the next assignments based on the previous ones. This mechanism continues in the same manner during the whole length of the simulation run. Therefore, the simulation results in this case become better and closer to the by-hand-results by increasing the replication length or the number of replications. Thus, the used number of replications is determined by trial and error to be 100 (the results remain nearly the same when the number of replications used is more than 100).

5.8.2 Abandonment Action Logic

In order to verify the abandonment action logic, a hypothetical case is assumed and implemented through the simulation model. The hypothetical case is that the number of arriving customers is only two and the handling time of the single scheduled agent, for those two arrivals, is greater than the maximum patience time. Additionally, the percentage of infinite patience customers is considered to be zero. Intuitively, the result of such case is that the first arriving customer will seize the single existing agent. Meanwhile, the other customer will be waiting in the queue with an assigned patience time that is less than the handling time of the first customer. Thus, the second customer will, for sure, abandon the queue resulting in one served customer and one abandoned customer. When the proposed simulation model is run under the described circumstances, the results are one served customer and one abandoned customer. This confirms the correctness of the used logic to represent the abandonment action (search and remove).

5.8.3 Processing Sequence

The processing sequence for any served customer starts by listening to the welcoming message; after that he/she waits in the queue, obtain the service, and then leave the system. The proposed logic to represent this sequence is verified by measuring the average time length of each activity of these activities and also measuring the average existence time of each customer within the system (called system time) from the simulation runs. Intuitively speaking, if the sum of the time of all activities equals the measured system time, then the simulated customer follows the modeled sequence and does not experience any other time delays other than the modeled ones. Running the simulation model, for example, with a replication length of eight-hour period and 100 replications confirms that intuition as shown in Table 5-13.

Table 5-13: Verification of processing sequence logic

Simulated Measure	Welcoming Message	Average Waiting Time	Average Handling Time	Average System Time
Simulated Value (seconds)	4	1.7667	57.0353	62.8032

5.9 Model Validation

According to the proposed validation framework, the first step in validating the developed simulation model is to prepare the real and simulated data sets to be compared. In preparing the simulated data set the model is run for 100 replications. The validation is performed according to the following steps.

1. ***Select Validation Parameter.*** The number of served customers in each daily hour period is selected to be the validation output parameter. This choice, in specific, is justified by the fact that this parameter is a function of the agents' level which is the main decision variable in the call centre context. Moreover, it is a function of the most important other uncontrollable output parameters such as abandonment volumes and waiting time in the system queue. Thus, this parameter is a result of all system interactions and parameters, and its validity yields the whole model validity.

2. ***Check Normality of Output Populations.*** After preparing both real and simulated data sets, the normality of the number of served customers in each one-hour period is tested using *Anderson-Darling normality test.* The data set that yields a p-value significantly greater than 0.05 is considered to be drawn from a normal population. Accordingly, the periods whose data sets could be considered to be drawn from a normal population are periods 1, 2, 3, 15, 17, and 24, as shown in Table 5-14.

3. ***Check Equality of Normal Populations Variances.*** For the data sets conforming to the normality assumption, the equality of variances of both real and simulated output data is evaluated using the F-test. The pairs (of real and simulated output data for each period) that yield a p-value less than 0.05 are considered of different variances. The pairs of that type are pairs of periods 15 and 24 while the remaining four pairs could be assumed to have equal variances. The F-test results are presented in Table 5-14.

4. ***Check Equality of Population Means.*** As for the data sets satisfying the normality assumption, the t-test is used to evaluate the equality of the means of real and simulated output data pairs for each period. The pairs that give a p-value less than 0.05 are considered to have unequal means and thus indicate that the model in that period seems invalid. Accordingly, all periods whose data sets pass the normality check are valid where they show high p-values as shown in Table 5-15.

Table 5-14: P-values of hypothesis tests used in validation of normal data sets

Statistical Test	Normality Test *Anderson-Darling-test*		Variance Equality *F-test*	Mean Equality *t-test*
Period No.	**Real Data p-value**	**Simulated Data p-value**	**p-value**	**p-value**
1	0.702	0.509	0.265	0.670
2	0.461	0.626	0.103	0.747
3	0.156	0.471	0.668	0.950
15	0.810	0.458	0.000	0.745
17	0.224	0.679	0.268	0.373
24	0.138	0.118	0.000	0.762

On the other hand, the periods whose data sets do not satisfy the normality assumption are treated using nonparametric tests. Thus, *Wilcoxon-Mann-Whitney* test of hypothesis on medians equality is used. The high p-values obtained using this test (as shown in Table 5-15), for all real and simulated output data pairs, show that all remaining periods are valid except period 11. However, if the non-normality of the real data is neglected in addition to the normality of the simulated data, then the period 11 would be valid according to the t-test with a p-value equal to 0.273.

Table 5-15: P-values of hypothesis tests used in validation of nonparametric data sets

Statistical Test	Mean Equality *Wilcoxon-Mann-Whitney*	Statistical Test	Mean Equality *Wilcoxon-Mann-Whitney*
Period No.	**p-value**	**Period No.**	**p-value**
4	0.0611	**13**	0.2334
5	0.9437	**14**	0.7555
6	0.8845	**16**	0.1909
7	0.4523	**18**	0.8714
8	0.9570	**19**	0.7117
9	0.3588	**20**	0.3375
10	0.0525	**21**	0.5582
11	0.0194	**22**	0.1995
12	0.1286	**23**	0.2190

Based on the obtained results, according to the proposed validation framework, the developed simulation model is validated and ready to be used in making decisions concerning the operations management of the call centre. This model validity, however, is obtained after doing several iterations in which the model was invalid. The model invalidity was investigated, as described in the proposed validation framework, using the *deterministic validation* technique. The deterministic validation showed that the model invalidity was due to problems in the input data models. Thus, the developed input data models were revisited and modified several times until the current valid model was obtained.

5.10 Optimal Staffing Decisions

The developed simulation model is utilized to make the optimal staffing decision in the studied call centre. What-if-analysis is applied using the simulation model to allocate the minimum required staff levels per daily period that satisfy a service level equal to or greater than 80%. In doing so, the model is run for each daily period, separately, but with considering the impact of the previous periods on the initial state of the studied period. This is to avoid the drawbacks of the independence assumption between the consecutive periods which may result in understaffing. The achieved results of this analysis are presented in Table 5-16.

Table 5-16: Optimal staffing decision

Period No.	No. of Agents	Obtained Service Level%	Confidence Interval
1	70	80.6061	± 2.103
2	54	81.4620	± 2.144
3	53	81.4615	± 2.170
4	19	81.9264	± 1.914
5	74	82.4514	± 1.202
6	59	83.6620	± 1.354
7	85	83.2719	± 1.763
8	17	82.9204	± 1.856
9	32	80.1702	± 1.897
10	73	80.6692	± 0.873
11	112	80.6867	± 2.089
12	129	80.0069	± 2.289
13	133	80.0456	± 2.297
14	142	80.1670	± 2.310
15	118	80.4971	± 2.211
16	98	80.0700	± 2.358
17	143	80.3325	± 2.342
18	107	80.8387	± 2.085
19	65	80.8137	± 2.110
20	74	82.9326	± 0.807
21	112	80.0606	± 1.873
22	110	80.2507	± 2.115
23	100	80.3894	± 2.154
24	93	80.4943	± 2.147

5.11 Optimal Shift Scheduling Decisions

After an appropriate staffing level has been determined for each period, this time-varying staffing level is fed into an integer linear programming program to select from all the feasible shifts that cover the staffing requirements the schedule of minimum cost.

There were 38 daily shifts considered to be candidate shifts for that selection process. The first shift starts at 07:30 AM and there is a new shift which starts after that every half-hour period. The last shift in the day starts at 02:00 AM.

In the call centre context, each shift must have some kind of break(s) for the agents working during that shift for two important purposes: stress relief and shift meal. Break allocation in each shift gives rise to several possible scenarios to be studied. There are two main scenarios that are considered in the current work as detailed below.

5.11.1 Shift Scheduling Scenarios

The first scenario for break allocation is to set fixed points in time for the breaks starting times. On the other hand, in the second scenario, the starting time of each break falls within a fixed time-window in which the starting time may vary. Each of these scenarios also produces another set of sub-scenarios that could appear due to the possible variations in the breaks duration and their numbers within each single shift. Three sub-scenarios are studied for each one of the main scenarios yielding to a total of six different scenarios investigated in this thesis. The scenarios are summarized as follows:

1. ***An-hour middle break.*** In this scenario, there is a single break allocated in each shift after the shift starting time by four hours.

2. ***An-hour middle breaks with 2-hours time-window.*** The number of agents working in each shift may be divided into two groups. One group with an hour middle break that starts after three working hours. The other group takes an hour break also but after 4 working hours.

3. ***Two half hour breaks.*** In this scenario, there are two breaks allocated in each shift. One break after two working hours. The other break starts after five and half hours from the starting time of the shift.

4. ***Two half hour breaks with 1.5-hours time-window.*** The number of agents working in each shift may be divided into three groups. One group with a half hour break that starts after 1.5 working hours and another half hour that starts after five hours from the starting time of the shift. The second group has a half hour break that starts after two working hours and another half hour that starts after five and half hours from the starting time of the shift. The last group has a half hour break that starts after 2.5 working hours and another half hour that starts after six hours from the starting time of the shift.

5. ***"15-30-15" minutes breaks.*** In this scenario, there are three breaks allocated in each shift. One 15-minutes break after two working hours. The second break is a 30-minutes break that starts after four hours from the starting time of the shift. The last one is a 15-minutes break that starts after 6.5 hours from the starting time of the shift.

6. ***"15-30-15" minutes breaks with two time-windows.*** The number of agents working in each shift may be divided into three groups. One group with a 15-minutes break after 1.75 working hours; the second break is a 30-minutes break that starts after 3.5 hours from the starting time of the shift; and the last one is a 15-minutes break that starts after 6.25 hours from the starting time of the shift. The second group has a 15-minutes break after two working hours; the second break is a 30-minutes break that starts after four hours from the starting time of the shift; and the last one is a 15-minutes break that starts after 6.5 hours from the starting time of the shift. The last group has a 15-minutes break after 2.25 working hours; the second break is a 30-minutes break that starts after 4.5 hours from the starting time of the shift; and the last one is a 15-minutes break that starts after 6.75 hours from the starting time of the shift.

5.11.2 Solution of Shift Scheduling Scenarios

The solution of the shift scheduling problem for each scenario produces two different outputs. The first output is the determination of the number of selected working shifts and their starting times (i.e., shifts daily pattern). The number of agents scheduled to work in each shift and the total number of agents per day is the second output.

Shifts Daily Pattern

The first output for scenarios 1, 3, and 5 is given as follows:

- ***Scenario one*** requires 13 different shifts out of all possible 38 shifts. Those shifts are shifts number 2, 4, 6, 8, 9, 10, 12, 20, 22, 24, 26, 28, and 34.
- ***Scenario three*** requires 18 different shifts out of all possible 38 shifts. Those shifts are shifts number 1, 2, 3, 4, 5, 6, 7, 19, 20, 21, 22, 23, 24, 25, 26, 34, 35, and 38.
- ***Scenario five*** requires 18 different shifts out of all possible 38 shifts. Those shifts are shifts number 1, 2, 3, 4, 5, 7, 8, 9, 11, 16, 17, 19, 20, 21, 24, 25, 26, 27, 33, 34, and 37.

The first output for scenarios 2, 4, and 6 is given by the number of shifts, starting times, and the different types of groups within each shift. This information is shown in Table 5.17, Table 5-18 and Table 5-19.

Table 5-17: Shift patterns for scenario 2

Shift No.	Group No.	Shift No.	Group No.	Shift No.	Group No.
1	1, 2	**12**	1	**24**	1
2	1, 2	**16**	2	**25**	1
3	1	**18**	2	**26**	1
5	1	**19**	2	**33**	1
6	1	**21**	2	**34**	2
7	1	**22**	1	**35**	2
8	1	**23**	1, 2	**37**	1
11	1				

Table 5-18: Shift patterns for scenarios 4

Shift No.	Group No.	Shift No.	Group No.
1	1	**20**	3
2	3	**21**	1
5	2	**23**	2
6	1, 2, 3	**24**	1, 2
7	3	**25**	3
8	2, 3	**26**	3
17	3	**35**	3
18	3	**37**	3
19	3	**38**	3

Table 5-19: Shift patterns for scenarios 6

Shift No.	Group No.	Shift No.	Group No.
1	1, 3	**18**	1, 2
2	3	**19**	3
3	1, 2	**23**	1,2
6	1, 2	**24**	1, 2
7	1	**25**	1, 2
8	1, 2, 3	**26**	1,2
9	1, 2	**33**	1, 2
16	3	**38**	3
17	1, 2		

Number of Scheduled Agents

The number of scheduled agents in each daily shift is the second component of the shift scheduling problem solution. This output for scenarios 1, 3, and 5 is presented in Table 5-23 while, Table 5-20, , Table 5-21 and Table 5-22 are present the same output for scenarios 2, 4, and 6.

Table 5-20: Number of scheduled agents/shift for scenarios 2

Shift No.	Number of Agents	Shift No.	Number of Agents	Shift No.	Number of Agents
1-1	16	**11-1**	1	**23-2**	11
1-2	15	**12-1**	2	**24-1**	12
2-1	10	**16-2**	5	**25-1**	12
2-2	37	**18-2**	25	**26-1**	13
3-1	17	**19-2**	11	**33-1**	1
5-1	17	**21-2**	11	**34-2**	3
6-1	61	**22-1**	10	**35-2**	7
7-1	16	**23-1**	11	**37-1**	6
8-1	90				

Table 5-21: Number of scheduled agents/shift for scenarios 4

Shift No.	Number of Agents	Shift No.	Number of Agents	Shift No.	Number of Agents
1-1	8	**8-2**	14	**23-2**	11
2-3	78	**8-3**	17	**24-1**	13
5-2	56	**17-3**	13	**24-2**	11
6-1	29	**18-3**	21	**25-3**	2
6-2	56	**19-3**	12	**26-3**	11
6-3	15	**20-3**	28	**35-3**	6
7-3	9	**21-1**	1	**37-3**	8
				38-3	3

Table 5-22: Number of scheduled agents/shift for scenarios 6

Shift No.	Number of Agents	Shift No.	Number of Agents	Shift No.	Number of Agents
1-1	36	**8-3**	24	**23-2**	5
1-3	24	**9-1**	24	**24-1**	12
2-3	27	**9-2**	15	**24-2**	5
3-1	12	**16-3**	3	**25-1**	12
3-2	2	**17-1**	10	**25-2**	9
6-1	36	**17-2**	12	**26-1**	12
6-2	36	**18-1**	3	**26-2**	2
7-1	24	**18-2**	20	**33-1**	3
8-1	22	**19-3**	7	**33-2**	6
8-2	7	**23-1**	10	**38-3**	8

Table 5-23: Number of scheduled agents/shift for scenarios 1, 3, and 5

Scenario One	
Shift No.	**Number of Agents**
2	42
4	35
6	96
8	53
9	44
10	25
12	5
20	6
22	16
24	23
26	62
28	9
34	17

Scenario Three	
Shift No.	**Number of Agents**
1	17
2	17
3	78
4	56
5	56
6	29
7	29
19	9
20	23
21	30
22	11
23	11
24	13
25	13
26	13
34	6
35	3
38	8

Scenario Five	
Shift No.	**Number of Agents**
1	81
2	19
3	53
4	27
6	34
7	12
8	27
9	54
11	12
16	17
17	4
19	21
20	21
21	5
24	2
25	1
26	29
27	33
33	6
34	6
37	5

5.11.3 Conclusion of Shift Scheduling Decision

One observation which could be drawn from the results is, that some periods are overstaffed with an excess number of agents. This is a direct consequence due to scheduling in an uncertain environment where the difference in service demand between consecutive periods is significant.

In order to decide on the most suitable scenario, two evaluation domains should be considered: quantitative and qualitative domain. The qualitative domain would be concerned with the impact of the allocated breaks numbers and start times on the agent's performance. On the other hand, the quantitative domain needs to use tangible values; it is the total number of daily scheduled agents in each scenario. As shown in Figure 5-20, all scenarios that consider time-windows on the breaks starting times show a smaller number of agents than those of fixed starting times for the breaks. Moreover, scenarios number 2, 3, 4, and 6 are all acceptable from the perspective of achieving the minimum number of scheduled agents on a daily basis. This is due to the small differences between the number of agents required in these four scenarios.

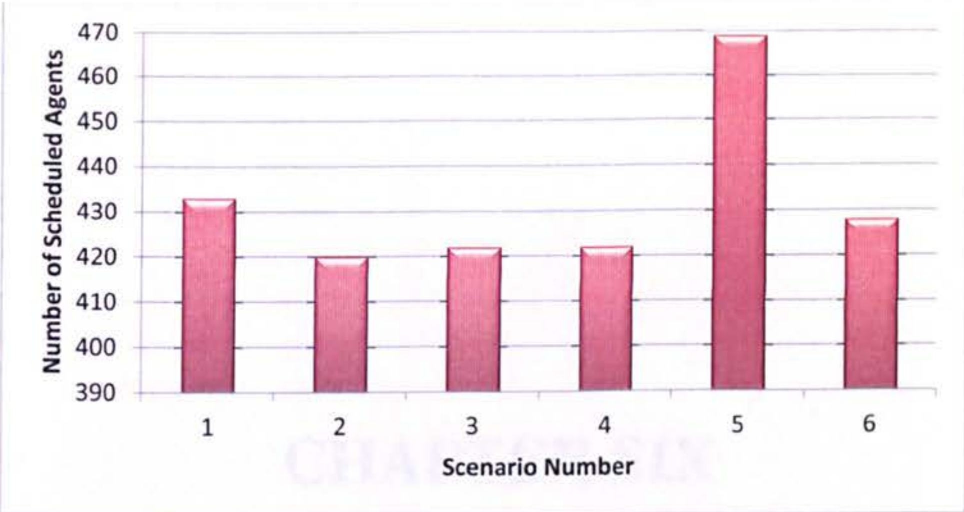


Figure 5-20: Scenarios comparison according to the number of excess agents

5.12 Chapter Conclusion

In this chapter, the details of the implementation of the proposed performance enhancement framework have been provided and it has been applied to one of the largest call centres in Libya as a case study. Our analysis has shown that the call centre under study did not adopt a clear management policy regarding its staffing decisions, with periods of understaffing and low service quality performance; and other periods of overstaffing having poor occupancy of the scheduled agents. Consequently, the proposed staffing decision making framework could yield cost reductions in addition to a more stable service level.

The abandons are affected by the average waiting time in the queue (which can be controlled by the call center),so by using this model the staff level will be managed and according to that the waiting time will be reduced which mean less abandonment callers. also by applying the developed model which results in reduced waiting time, this will lead to improve customer satisfaction.

CHAPTER SIX

DISCUSSIONS AND IMPLICATIONS

CHAPTER SIX

DISCUSSIONS AND IMPLICATIONS

6.1 Discussion and analysis of output

After developing the stochastic data and simulation model, at this instant, the output metrics of the current system are evaluated using the developed simulation model. There are two time frames in which these measures could be evaluated either separately for each daily one-hour-period or aggregated over the course of the working day. The analysis of both time frames is discussed in the following sections.

6.1.1 Daily One-Hour-Period Analysis

According to the proposed time frames of analysis in section 4.1.9, the analysis of daily one-hour-period performance measures fall under the terminating analysis category. In order to perform the terminating analysis, a number of several independent simulation runs should be conducted. Afterwards, the obtained simulated output data for each performance measure is evaluated by a point estimator (the mean value of the simulated data) and an error estimator for the point estimator (the confidence interval). The initial number of conducted simulation runs h_o is considered to be one-hundred runs (replications) as discussed previously. After that, the obtained confidence intervals are examined whether they are small enough so that the obtained point estimates could be considered to be reliable or not. The number of desired runs is evaluated using equation 4-3. The results of this analysis are as follows:

Table 6-1: Terminating Analysis of Service Level (SL %)

Period No.	Mean Value (θ)	Confidence Interval (CI)	Period No.	Mean Value (θ)	Confidence Interval (CI)	Period No.	Mean Value (θ)	Confidence Interval (CI)
1	46.76	± 2.59	9	97.10	± 0.197	17	85.94	± 2.06
2	51.35	± 2.92	10	97.35	± 0.095	18	97.18	± 1.914
3	77.16	± 3.05	11	86.96	± 3.145	19	100.00	± 0.263
4	97.36	± 1.30	12	68.20	± 2.778	20	97.49	± 0.30
5	98.80	± 0.74	13	46.64	± 2.98	21	86.04	± 2.607
6	97.87	± 0.779	14	47.25	± 2.986	22	86.61	± 2.093
7	94.62	± 2.52	15	71.13	± 2.569	23	78.97	± 2.293
8	91.70	± 2.119	16	83.28	± 2.249	24	60.47	± 2.834

Table 6-2: Terminating Analysis of Average Speed of Answer (ASA in seconds)

Period No.	Mean Value $(\overline{\theta})$	Confidence Interval (CI)	Period No.	Mean Value $(\overline{\theta})$	Confidence Interval (CI)	Period No.	Mean Value $(\overline{\theta})$	Confidence Interval (CI)
1	31.54	± 1.82	9	0.02	± 0.011	17	6.97	± 1.139
2	28.26	± 1.89	10	0.06	± 0.085	18	0.75	± 0.797
3	11.53	± 1.68	11	5.55	± 1.595	19	0.02	± 0.013
4	0.38	± 0.37	12	18.95	± 1.847	20	0.42	± 0.249
5	0.04	± 0.027	13	45.04	± 3.07	21	5.22	± 0.967
6	0.09	± 0.097	14	48.29	± 3.649	22	5.33	± 0.784
7	1.38	± 1.02	15	11.44	± 1.108	23	8.71	± 1.081
8	2.30	± 0.852	16	8.43	± 1.298	24	18.74	± 1.584

Table 6-3: Terminating Analysis of Abandonment Percentage%

Period No.	Mean Value $(\overline{\theta})$	Confidence Interval (CI)	Period No.	Mean Value $(\overline{\theta})$	Confidence Interval (CI)	Period No.	Mean Value $(\overline{\theta})$	Confidence Interval (CI)
1	16.57	± 0.74	9	1.55	± 0.103	17	2.97	± 0.29
2	13.37	± 0.72	10	1.09	± 0.055	18	1.80	± 0.207
3	7.77	± 0.80	11	2.31	± 0.384	19	1.50	± 0.107
4	3.86	± 0.76	12	5.62	± 0.388	20	1.83	± 0.269
5	3.82	± 0.33	13	11.15	± 0.60	21	4.88	± 0.619
6	3.85	± 0.306	14	10.30	± 0.585	22	4.05	± 0.394
7	3.69	± 0.68	15	7.44	± 0.460	23	6.56	± 0.503
8	3.53	± 0.593	16	4.05	± 0.423	24	11.99	± 0.732

Table 6-4: Terminating Analysis of Agent Occupancy%

Period No.	Mean Value $(\overline{\theta})$	Confidence Interval (CI)	Period No.	Mean Value $(\overline{\theta})$	Confidence Interval (CI)	Period No.	Mean Value $(\overline{\theta})$	Confidence Interval (CI)
1	94.5	± 0.5	9	21.88	± 1.70	17	81.73	± 0.9
2	91.74	± 0.8	10	59.84	± 0.027	18	70.91	± 0.028
3	77.53	± 1.6	11	84.64	± 1.20	19	46.03	± 0.028
4	46.23	± 3.5	12	94.40	± 0.5	20	74.45	± 2.6
5	33.13	± 3.3	13	97.13	± 0.3	21	85.76	± 1.2
6	27.70	± 3.4	14	96.50	± 0.4	22	86.50	± 0.9
7	43.73	± 4.7	15	92.47	± 0.6	23	87.22	± 0.8
8	46.22	± 2.8	16	78.49	± 1.1	24	91.23	± 0.8

6.1.2 Aggregated Daily Analysis

The analysis of performance measures over the course of the working day falls under the steady-state analysis category. The steady-state analysis, in this thesis, is conducted using the truncated replications strategy. This strategy requires running the model also for several independent runs whose run-length is equal to twenty-four hours (the working day period). The initialization bias is then truncated by identifying a warm-up period, after which the model parameters start to be evaluated. The number of replications is determined in the same manner as that for the terminating analysis discussed in the previous section which results in 100 replications. The steady-state analysis of the current system for cluster one gives the following results:

Table 6-5: Steady-state analysis of cluster-one days

Performance Measure	Mean Value $(\overline{\theta})$	Confidence Interval (CI)
Service Level (SL %)	71.78	± 0.783
(ASA in seconds)	16.10	± 0.61
Abandonment Percentage%	**6.86**	**± 0.16**
Agent Occupancy%	71.13	± 0.2

6.1.3 Production Runs and Analysis

The previous output analysis is conducted on the current state of the studied call centre. This state is considered as a base scenario from which the typical system behaviour could be understood especially that some of performance measures are not considered in the real call centre records. In the following sections, the current system behaviour is investigated, and then sensitivity analysis of the performance measures to changes in the staff levels is conducted and discussed.

6.1.4 Current System Behaviour

Studying the output values for both the *average-speed-of-answer* and the *agent occupancy* yields some interesting conclusions. This call centre does not follow a specific management regime. However, it is *efficiency-driven* in some periods and *quality-driven* in the other periods. This could be concluded by examining Figure 6-1and Figure 6-2. The efficiency-driven periods are 1, 2, 12, 13, 14, 15, and 24 where they experience high agent's occupancy (utilization) that exceeds 90% in addition to high waiting times varying between 12 to 49 seconds on the average. The remaining periods are considered *quality-driven* where the waiting times are very low (below 8 seconds) in addition to low agents' utilization. This leads to an important conclusion that the studied centre staffing/customer service objectives are not properly controlled and managed. This is very clear due to the use of two management regimes that treats customer satisfaction differently which leads to unstable system performance.

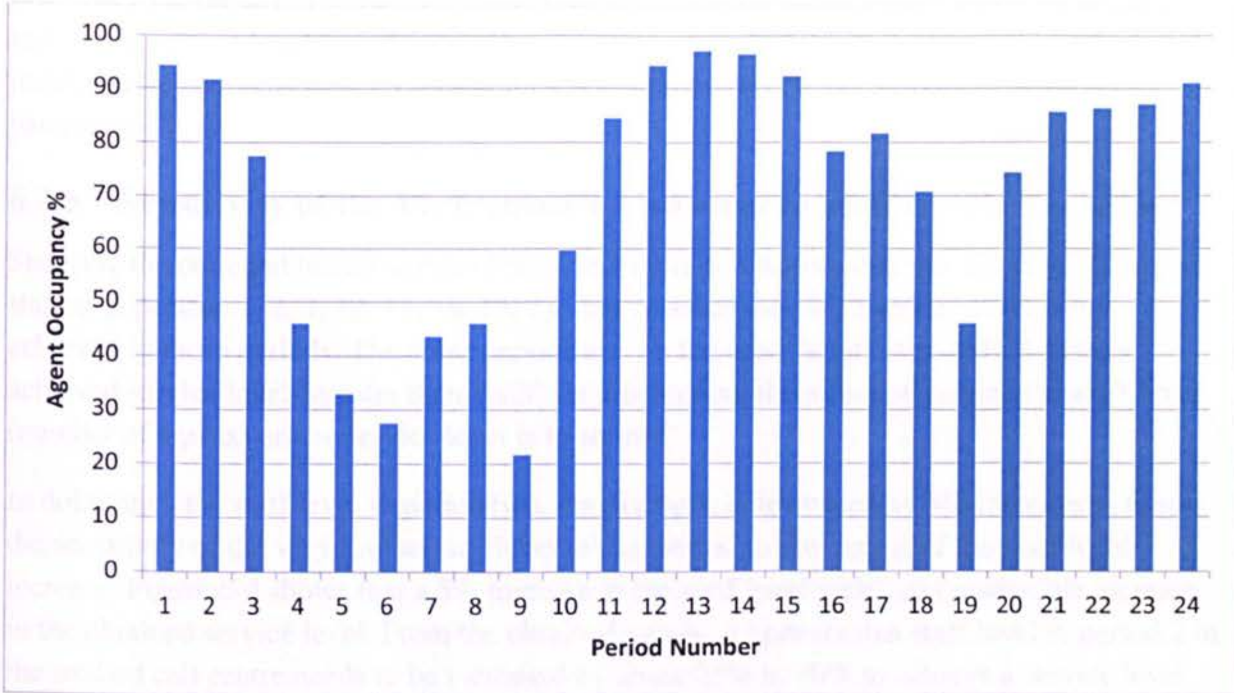


Figure 6-1: Average agent occupancy in daily hour-periods

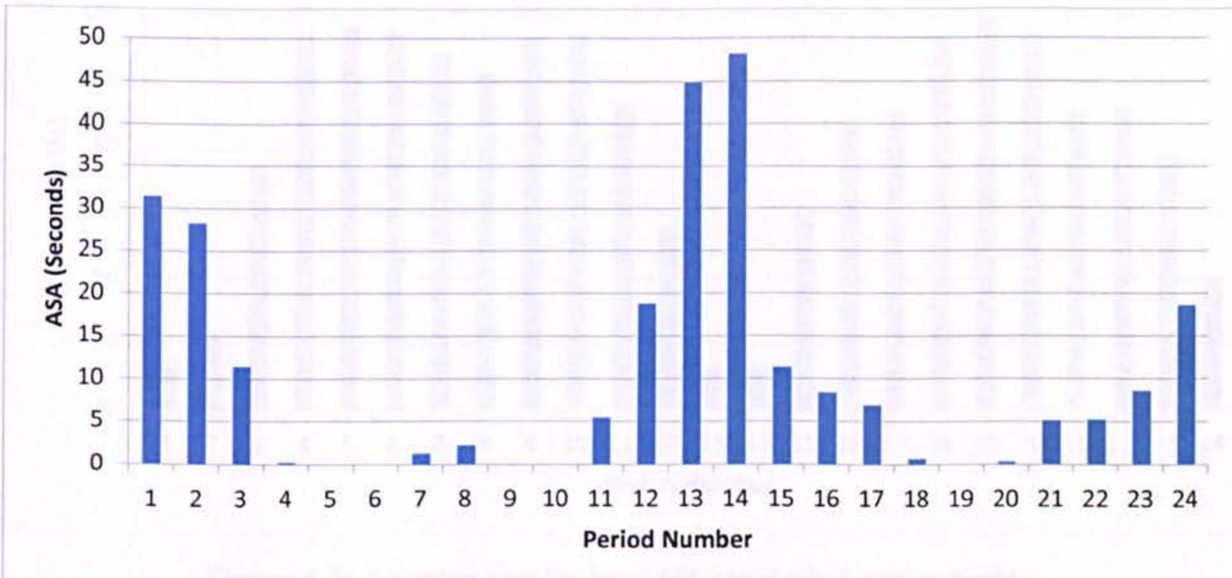


Figure 6-2: Average speed of answer (ASA) in daily hour-periods

According to this conclusion, managing the call centre under *quality-efficiency-driven* (QED) regime would be the best operating regime. This is due to the ability of this regime to achieve some balance between operations efficiency and customer service quality.

The performance measure that relates the efficiency with the quality is the service level since it represents the desirable percentage of served calls within allowable waiting time of customers. The most commonly used service level is 80/20 which stands for operating the centre under the constraint of answering 80% of the incoming calls within 20 seconds of their waiting in the system queue for the acquired service. Therefore, the focus of the next sections is to study the impact of variations in the staff levels on the performance measures generally and on the service level specifically. The objective of performing this sensitivity analysis is to justify the selection of staff levels and their schedules to be the performance enhancement parameters.

6.1.5 Sensitivity of the Performance Measures to Staff Levels Variation

Studying the obtained hourly service levels, in Figure 6-3, shows that the call centre is under-staffed in periods 1, 2, 3, 12, 13, 14, 15, 23, and 24 since the 80/20 service level is not achieved in those periods. The other periods are, on the other hand, over-staffed as they achieved service levels greater than 80/20. In this section, the effect of varying the staff level (number of agents) on the service level is examined.

In doing that, the staff level in period two, for example, is increased by 5% increments to test the sensitivity of the very low service level in that period to the impact of the staff level increase. Figure 6-4 shows that a 5% increase in the staff level yields a considerable increase in the obtained service level. From the obtained values, it appears that staff level in period 2 in the studied call centre needs to be increased by about 25% to 30% to achieve a service level greater that 80%.

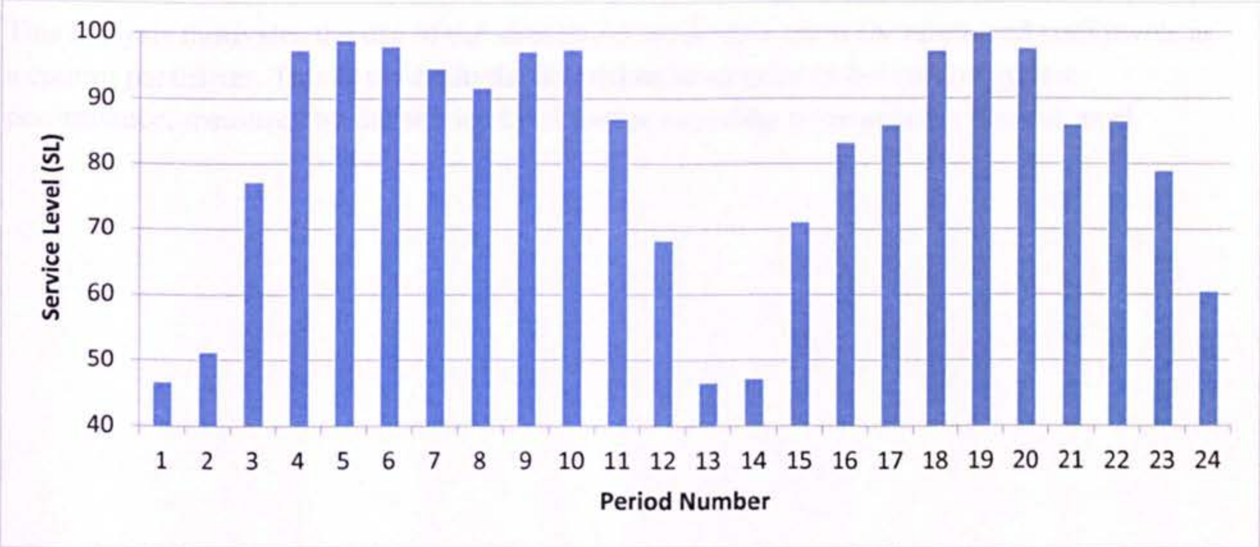


Figure 6-3: Average service level (SL) in daily hour-periods

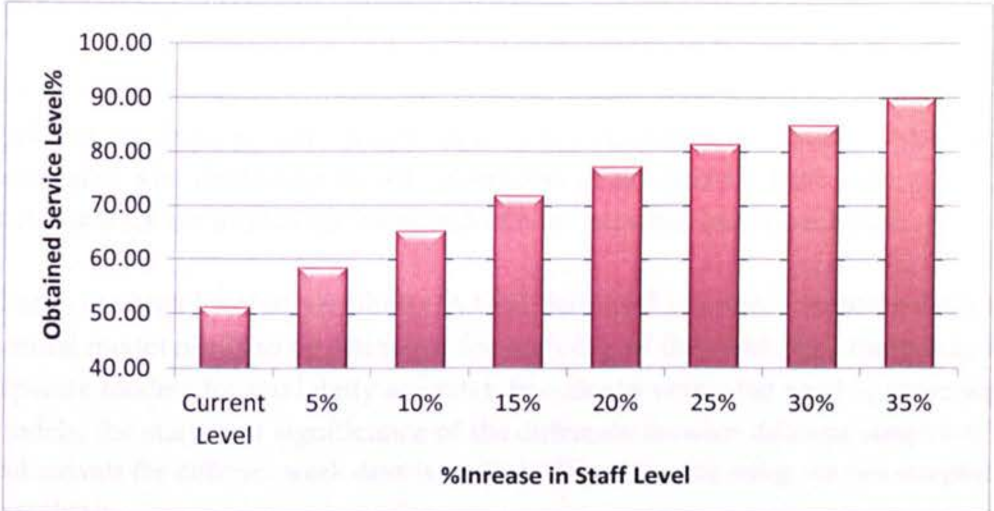


Figure 6-4: Values of service level (SL) at different values of staff levels

This analysis motivates the use of the simulation model in a quest for optimized staff levels as a control parameter. This is to obtain the desired enhancement in the current system performance, measured by the service level metric adjusting it towards the desired level.

6.2 Implications of the Thesis

- In this study, two new models are proposed. The first one is for the arrival counts and the second one is for the abandonment rates.

- A new doubly stochastic model has been proposed to represent the arrival counts in different one-hour periods, which successfully addresses the various challenges associated with modelling the arrival process in call centres. Moreover, a new methodology for modelling the abandonment rates has been developed.

- Due to the high level of variability in total daily call volume, a separate daily arrival normal model needs to be estimated for each day of the week (i.e., there may be seven separate models for total daily arrivals). In order to verify the need to these separate models, the statistical significance of the difference between different samples of the daily call arrivals for different week days is studied. This is done by using the two samples t-test of hypothesis.

- Customer abandonment phenomenon occurs due to poor service and bad staffing plans; indeed any reduction in abandonment volumes even by one percent is a direct consequence of improved service and/or staffing plans. Thus, the abandonment volumes should be modelled carefully to create accurate models that represent the operations of a call centre successfully. The modelling of customer abandonment behaviour, however, is not an easy task. The main difficulty here is that the percentage of customers that are willing to abandon and patience time after which they abandon are not absolute input parameters. In essence, their values depend on the system performance (i.e., poor and slow service leads to long queuing time and thus high abandonment volumes). Now, the question is how to model the abandonment behaviour as an input parameter, and at the same time, as an output performance measure. Output performance measures are those with values that are sensitive to the changes in the system design parameters (e.g. number of staffed agents).

- In the current work, a new modelling methodology is proposed to account for this difficulty. The proposed methodology represents the abandonment volumes using a couple of parameters, where abandonments are assumed to be of a multi-modal parameter. This assumption is supported by the case study applied in this thesis, where the hazard rate for abandonment (the time phased probability for abandoning) is found to be multi-modal. Therefore, each time-phased mode of abandonment is represented using two parameters. The first parameter is the time interval itself (patience time that after spending it in the queue the customer will abandon). The second is the percentage/weight of customers who belong to this patience-time-interval.

- The realistic input data modelling and simulation help greatly in solving the staffing problem to determine a realistic minimum number of staff members in each of the different daily periods. Traditionally, those realistic estimates are used as inputs to the set covering problem to solve the shift scheduling problem to optimality.

- Studying the output values for both the *average-speed-of-answer* and the *agent occupancy* yields some interesting conclusions. This call centre does not follow a specific management regime. However, it is *efficiency-driven* in some periods and *quality-driven* in the other periods.. This leads to an important conclusion that the studied centre staffing/customer service objectives are not properly controlled and managed. This is very clear due to the use of two management regimes that treats customer satisfaction differently which leads to unstable system performance.

6.3 Discussion of potential benefits when applying the developed model in the call centre:

- Abandon rate is a typical measure of call center performance. It should be noted, however, that abandon rate is not entirely under the call center's control.

- While abandons are affected by the average waiting time in the queue (which can be controlled by the call center),so by using this model the staff level will be managed and according to that the waiting time will be reduced which mean less abandonment callers.
- Customer satisfaction is one of the most critical metrics for any call centre. Studies have revealed, and common sense supports, a critical and direct correlation between customer satisfaction, customer loyalty, corporate revenues and employee morale and performance.so by applying the developed model which results in reduced waiting time, this will lead to improve customer satisfaction.

- The shift scheduling benefits are numerous. The most important benefits for a business are that they can save a lot of time and money. By using this model, it will make the company more likely that staffing levels will be adequate for any given shift. It is also worth mentioning that staff members tend to favour this type of system once they get used to it.

- By using the proposed model, Changing shifts is straight forward and many of the scheduling tasks can be automated; this means there will be no more need to spend hours dealing with excel sheets or going through plough of papers if manual schedule is the practice even worse plain old paper. It will be possible to turn around a normal scheduling task which could take hours to just a few minutes.

CHAPTER SEVEN

CONCLUSIONS AND FUTURE WORK

CHAPTER SEVEN
CONCLUSIONS AND FUTURE WORK

7.1 Conclusions

In this thesis, a framework was proposed and developed for call centres so that reliable performance evaluation and effective enhancement can be realised. This is achieved through making realistic optimal staffing and scheduling decisions based on accurate system modelling. The framework proceeds in three phases:

Phase I. Stochastic modelling of the system parameters.

Phase II. Simulation modelling of the call centre operations.

Phase III. Optimization of staffing levels and shift scheduling decisions.

In phase I, the framework begins by developing stochastic data models for call centre operations parameters which have been divided into service demand (arrival volumes) and service quality (service times, abandonment volumes, and patience time) parameters. In this respect, the current work achieves intended contributions in modelling both arrival counts and the abandonment process.

Studying the arrival data patterns in a real call centre reveals three important properties:

- *Arrival rate uncertainty* with multiple levels of variability (i.e., monthly, weekly, day-to-day, and daily-period-to-period variability levels),
- *Data over-dispersion* (i.e., the variance of the data is much higher than its mean), and
- *Strong correlation* (i.e., there is positive correlation between the volumes of arrivals in the successive periods where 83% of the periods-pairs shows correlation coefficient greater than 0.5).

Each of these properties has a great impact on the validity of any model developed for representing the arrival data. In this thesis, a new model was proposed to capture the effects of these properties with a new approach aimed at seeking a better fit to the real data and involving fewer estimated parameters than other models suggested by literatures.

The developed model succeeds in handling and representing the real data properties properly. Moreover, the arrival data model is validated by showing negligible difference from the real data points.

As for the service times, the call centre in concern also experienced a good amount of uncertainty and were therefore modelled using a log-normal distribution with *time-varying* parameters. This approach has a strong support according to the literatures which were reviewed. It was also found that the successive periods may form clusters which have similar service rates.

Regarding the abandonment behaviour, a new model was proposed to represent customers' abandonment through the use of a two important parameters. This is due to the fact obtained from the case study data that the time phased probability for abandoning is multi-modal. The two key parameters are different clusters of patience time and the corresponding customers' proportions with respect to the total number of arriving customers. The proposed model succeeds in modelling the abandonment rate as an input parameter while maintaining its sensitivity, as a performance measure, to the variations in the control parameters. Moreover, the developed model has managed to achieve high validity when its results were compared to the real data obtained from the studied call centre.

In the second phase of the proposed framework, the developed data models were then used as input data pattern into a simulation model that was used to determine the minimum staffing levels in daily one-hour periods.

Developing the simulation model involved several iterations of modifications for both the conceptual model of the call centre operations and the input data models. These iterations provided a reliable indicator on the robustness of the proposed data models which, as a result, brings the developed framework closer to reality.

The developed simulation model was validated using a proposed validation methodology. This methodology ensures a high level of confidence in the results of the validation process as it does not depend on the common assumptions which are normally used in simulation validation.

The validation is done in two levels: the aggregated output measures over the course of the day and the hourly measures level. This increases the credibility of the developed model.

Once the simulation model was validated, it was then used to perform sensitivity analysis on the impact of different staff levels on the call centre performance. It was noted by varying the staff levels even by small percentages can produce a considerable impact on the overall system performance.

In phase III, the staffing and scheduling problems were solved using a two-steps approach. The staffing problem was solved by applying a simulation-based what-if-analysis on the number of agents and the corresponding service level in a given one-hour period. This was done by running the simulation model with various values of staff levels for each daily period till reaching the staff level that will satisfy the pre-determined service quality constraint (i.e., the optimal staff level).

These staffing levels were used as input to an IP model which has managed to optimally allocate the service agents to the different operating shifts of a working day.

The IP model was formulated as a set-covering model that was used to achieve final optimisation.

The set-covering formulation was used in developing the staff schedules for six different scenarios.

Studying those scenarios shows that allocating the breaks starting times by using flexible time-windows can produce better solutions than fixing the starting times. Moreover, there are four scenarios that are suitable from the perspective of minimising the number of agents.

7.2 Recommendations for Future Work

Based on the research results, the author would like to recommend the following tasks which can extend and enhance the functionalities of the developed framework:

- The proposed arrivals model can be further developed to obtain better representation for the correlation between arrival volumes in the successive daily periods. This will enable the research results to be applied to a wider call centre community which may have more complicated arrival scenarios and work patterns.
- The developed simulation model abilities can be further enhanced by adding more types of service agents and different logics for routing to cater for different call types. This may include transforming the call centre to a multi-skill call Centre, as the studied call centre only employs single-skill operators at present, also study the multi-skill call centres leads to the requirement of a simulation-based optimisation algorithm to determine the optimal staff level in each daily period for each skill group. This can be compared to a single-skill call centre which normally only requires using some sort of what-if-analysis to allocate the optimal staffing levels in order to satisfy the targeted service requirement.
- At present the developed framework produces acceptable solutions based on the author's own calculations. However in order to realise the full potential benefits, the author suggested that that the framework should be implemented in the associated call centre for systematic evaluation for its accuracy and effectiveness.
- At present the developed frame work requires pre-populated data for its performance evaluation process. This data would be normally based on historical trend etc. It is suggested that a dynamic feed of live data link can be developed so that the performance can be reviewed on a real time and continuously basis. This will greatly enhance the model's capability to deal with more dynamic and unpredictable scenarios.

- It is envisaged that the developed framework and modelling algorithm developed for call centres can be applied, once relevant modifications have been made, in other industrial sectors which bear similar characteristics. This includes banks, health care centres, production line and manufacturing. It will allow these organisations to conduct similar performance and scheduling evaluation for future improvement.
- More qualitative aspects of call centre management can be used to design different operating scenarios and use them as standard templates. This may include the scope of agent's satisfaction besides considering the customer's satisfaction.
- The present framework developed for the call centre has limited ability to evaluate staff and other operating costs. The author suggests that a link can be developed to feed the outputs of the model into an additional financial package to provide a more accurate cost indicator when different staff strategies are evaluated.

REFERENCES

Aksin, Z., Armony, M., and Mehrotra, V., "The Modern Call Centre: A Multi-Disciplinary Perspective on Operations Management Research," *Production And Operations Management*, vol. 16, no. 6, pp. 665–688, 2007.

Alfares, H., "Survey, Categorization, and Comparison of Recent Tour Scheduling Literature," *Annals of Operations Research*, vol. 127, no. 1-4, pp. 145-175, 2004.

Alex, C. 2010), cited from http://www.callcentrehelper.com/the-top-ten-call-centre-solutions-12632.htm(November 14,2012)

Andrews, D., "An Empirical Process Central Limit Theorem for Dependent Non-Identically Distributed Random Variables," *Journal of Multivariate Analysis*, vol. 38, no. 2, pp. 188-203, 1991.

Armony, M. and Ward, A., "Fair Dynamic Routing in Large-Scale Heterogeneous-Server Systems," *Operations Research*, vol. 58, no. 3, pp. 624-637, 2010.

Artalejo, J. and Pla, V., "On The Impact of Customer Balking, Impatience and Retrials in Telecommunication Systems," *Computers & Mathematics with Applications*, vol. 57, no. 2, pp. 217–229, 2009.

Avramidis, A., Deslauriers, A., and L'Ecuyer, P., "Modeling Daily Arrivals to a Telephone Call Centre," *Management Science*, vol. 50, no. 7, pp. 896-908, 2004.

Aykin, T., "Optimal Shift Scheduling with Multiple Break Windows," *Management Science*, vol. 42, no. 4, pp. 591-602, 1996.

Aykin, T., "A Comparative Evaluation of Modeling Approaches to the Labor Shift Scheduling Problem," *European Journal of Operational Research*, vol. 125, no. 2, pp. 381–397, 2000.

Atlason, J., Epelman, M., and Henderson, S., "Call Centre Staffing with Simulation and Cutting Plane Methods," *Annals of Operations Research*, vol. 127, no. 1, pp. 333–358, 2004.

Ashley, David. (1995) A Spreadsheet Optimization System for Library Staff Scheduling.Computers Operations Research, 22 No. 6, 615-624.

Baker, K. & Magazine, M. (1977) Workforce Scheduling with Cyclic Demands and Day-off Constraints. Management Science, 24 No. 2.

Baker, K. & Magazine, M. (1977) Workforce Scheduling with Cyclic Demands and Day-off Constraints. Management Science, 24 No. 2.

Banks, Carson, Nelson, and Nicol, *Discrete-Event System Simulation*, 3rd ed.: Prentice Hall, pp. 16-20, 2000.

Bekker, J. and Viviers, L., "Using Computer Simulation to Determine Operations Policies for a Mechanized Car Park," *Simulation Modelling Practice and Theory*, vol. 16, no. 6, pp. 613–625, 2008.

Ben-Israel, A., "A Newton-Raphson method for the solution of systems of equations," *Journal of Mathematical Analysis and Applications*, vol. 15, no. 2, pp. 243-252, 1966.

Bhulai, S., Koole, G., and Pot, A., "Simple Methods for Shift Scheduling in Multiskill Call Centres," *Manufacturing & Service Operations Management*, vol. 10, no. 3, pp. 411–420, 2008.

Brown, L., Zhao, L., Zeltyn, S., Mandelbaum, A., Sakov, A. and Gans, N., "Statistical Analysis of a Telephone Call Centre," *Journal of the American Statistical Association*, vol. 100, no. 469, pp. 36-50, 2005.

Bouzada, M., "Dimensioning a Call Centre Simulation or Queue Theory," *Journal of Operations and Supply Chain Management*, vol. 2, no. 2, pp. 34-46, 2009.

Brusco, M. and Jacobs, L., "Optimal Models for Meal-Break and Start-Time Flexibility in Continuous Tour Scheduling," *Management Science*, vol. 46, no. 12, pp. 1630-1641, 2000.

Brusco, M. and Jacobs, L., "Starting Time Decisions in Labor Tour Scheduling: An Experimental Analysis and Case Study," *European Journal of Operational Research*, vol. 131, no. 3, pp. 459–475, 2001.

Bureau of Labor Statistics. United States Department of Labor. [Online]. http://www.bls.gov/ooh/ocos280.htm. (May 04, 2013).

Castillo, I., Joro, T.,and , Li, Y., "Workforce Scheduling with Multiple Objectives," *European Journal of Operational Research*, vol. 196, no. 1, pp. 162–170, 2009.

Chokshi R. (1999). Decision support for call center management using simulation. Proceedings of the 1999 Winter Simulation Conference.

Cleveland B andMayben J (1997). Call Center Management on Fast Forward.

Dantzig, G., "A Comment on Edie's "Traffic Delays at Toll Booths"," *Journal of the Operations Research Society of America*, vol. 2, no. 3, pp. 339–341, 1954.

Deslauriersa, A., L'Ecuyera, P., Pichitlamkenb, J., "Markov chain models of a telephone call centre with call blending," *Computers & Operations Research*, vol. 34, no. 6, pp. 1616–1645, 2007.

Dietz, D., "Practical scheduling for call centre operations," *Omega*, vol. 39, no. 5, pp. 550–557, 2011.

Ernst, A.T., Jiang, H., Krishnamoorthy, M., Sier, D. (2003) Staff Scheduling and Rostering:A Review of Applications, Methods and Models. European Journal of OperationalResearch, 153, 3-27.

Ertogral, K. and Bamuqabel, B., "Developing staff schedules for a bilingual telecommunication call centre with flexible workers," *Computers & Industrial Engineering*, vol. 54, no. 1, pp. 118–127, 2008.

Fay, M. and Proschan, M., "Wilcoxon-Mann-Whitney or T-Test? On Assumptions for Hypothesis Tests and Multiple Interpretations of Decision Rules," *Statistics Surveys*, vol. 4, pp. 1–39, 2010.

Feng, C., "Dynamic Operator Staffing Problem for Call Centres with Uncertain Arrival Rate," in *Proceedings of the International Conference on Business Computing and Global Informatization (BCGIN), 2011*, Shanghai, China, 2011, pp. 619 - 621.

Ferguson, T., "A Bayesian Analysis of Some Nonparametric Problems," *The Annals of Statistics*, vol. 1, no. 2, pp. 209-230, 1973.

Gans, N., Koole, G., and Mandelbaum, A., "Telephone Call Centres: Tutorial, Review, and Research Prospects," *Manufacturing & Service Operations Management*, vol. 5, no. 2, pp. 79–141, 2003.

Gans, N., Liu, N., Mandelbaum, A., Shen, H., and Ye, H., "Service times in call centres: Agent heterogeneity and learning with some operational consequences," *Borrowing Strength: Theory Powering Applications – A Festschrift for Lawrence D. Brown. IMS Collections*, vol. 6, pp. 99-123, 2010.

Garnet, O., Mandelbaum, A., and Reiman, M., "Designing a Call Centre with Impatient Customers," *Manufacturing & Service Operations Management*, vol. 4, no. 3, pp. 208-227, 2002.

Green, L., Kolesar, P., and Whitt, W., "Coping with Time-Varying Demand When Setting Staffing Requirements for a Service System," *Production and Operations Management*, vol. 16, no. 1, pp. 13–39, 2007.

Henderson, S., "Estimation for nonhomogeneous Poisson processes from aggregated data," *Operations Research Letters*, vol. 31, no. 5, pp. 375-382, 2003.

Heyman, D. and Sobel, M., *Stochastic Models in Operations Research:Stochastic Processes and Operating Characteristics*, 1st ed. Mineola, NY: Dover Publications, pp. 158-159, 2004.

Ingolfsson, A., Akhmetshina, E., Budge, S., Li, Y., and Wu, X., "A Survey and Experimental Comparison of Service Level Approximation Methods For Non-Stationary M/M/s Queueing Systems," *INFORMS Journal on Computing*, vol. 19, no. 2, pp. 201-214, 2007.

Jacobs, L. and Brusco, M., "Overlapping Start-Time Bands in Implicit Tour Scheduling," *Management Science*, vol. 42, no. 9, pp. 1247-1259, 1996.

Jongbloed, G. and Koole, G., "Managing uncertainty in call centres using Poisson mixtures," *Applied Stochastic Models in Business and Industry*, vol. 17, no. 4, pp. 307-318, 2001.

Jordan, J., Melouk, S., and Faas, P., "Analyzing Production Modifications of a C-130 Engine Repair Facility Using Simulation," in *Proceedings of the 2009 Winter Simulation Conference*, Austin, TX, 2009, pp. 1770-1777.

Kalaignanam, K. and Varadarajan, R., "Offshore Outsourcing of Customer Relationship Management: Conceptual Model and Propositions," *Journal of the Academy of Marketing Science*, vol. 40, no. 2, pp. 347-363, 2012.

Kao, E. and Chang, S., "Modeling Time-Dependent Arrivals to Service Systems: A Case in Using a Piecewise-Polynomial Rate Function in a Nonhomogeneous Poisson Process," *Management Science*, vol. 34, no. 11, pp. 1367-1379, 1988.

Kelton, W., Sadowski, R., and Swets, N., *Simulation with Arena*, 5th ed.: McGraw-Hill, pp. 276-278 and 320-328, 2010.

Kelton WD, Sadowski RP and Sturrock D (2004). Simulation with ARENA, 3rd edn, McGraw-Hill: New York.

Koole, G. and Mandelbaum, A., "Queueing Models of Call Centres: An Introduction," *Annals of Operations Research*, vol. 113, no. 1-4, pp. 41-59, 2002.

Law, A., "How to Build Valid and Credible Simulation Models," in *Proceedings of the 2005 Winter Simulation Conference*, Orlando, FL, USA, 2005, pp. 24-32.

L'Ecuyer, P., "Modeling and Optimization Problems in Contact Centres," in *Proceedings of the 3rd international conference on the Quantitative Evaluation of Systems*, Washington, DC, 2006, pp. 145 - 156.

Longo, F., "Design and Integration of the Containers Inspection Activities in The Container Terminal Operations," *International Journal of Production Economics*, vol. 125, no. 2, pp. 272–283, 2010.

Mahfouz, A., Hassan, S., and Arisha, A., "Practical Simulation Application: Evaluation of Process Control Parameters in Twisted-Pair Cables Manufacturing System," *Simulation Modelling Practice and Theory*, vol. 18, no. 5, pp. 471–482, 2010.

Mandelbaum, A., Sakov, A., and Zeltyn, S., "Empirical Analysis of a Call Centre," Working Paper 2002.

Massey, W., Parker, G., and Whitt, W., "Estimating the parameters of a nonhomogeneous Poisson process with linear rate," *Telecommunications Systems*, vol. 5, no. 2, pp. 361-388, 1996.

Mason, A. Ryan, D. and Panton, D., "Integrated Simulation, Heuristic and Optimisation Approaches to Staff Scheduling," *Operations Research*, vol. 46, no. 2, pp. 161-175, 1998.

Matteson, D., McLean, M., Woodard, D., and Henderson, S., "Forecasting Emergency Medical Service Call Arrival Rates," *The Annals of Applied Statistics*, vol. 5, no. 2B, pp. 1379-1406, 2011.

Mehrotra, V. and Fama, J., "Call Centre Simulation Modeling: Methods, Challenges, and Opportunities," in *Proceedings of the 2003 Winter Simulation Conference*, Louisiana, 2003, pp. 135-143.

Mitchell, B., "A Comparison of Chi-Square and Kolmogorov-Smirnov Tests," *The Royal Geographical Society*, vol. 3, no. 4, pp. 237-241, 1971.

Morris, J. & Showalter, M. (1983) Simple Approaches to Shift, Days-off and Tour Scheduling Problem. Management Science, 28 No. 8, 942-950

Nagarajan, K., Awyzio, J., and Vial, P., "The Use of ARENA Simulation Software to Illustrate Network Operations in an Educational Setting Using Case Studies," in *Proceedings of 7th World Multi-Conference on Systemics, Cybernetics and Informatics (SCI 2003)*, Orlando, Florida, USA, 2003.

Pichitlamken, J., Deslauriers, A., L'Ecuyer, P., and Avramidis, A., "Modeling and Simulation of A Telephone Call Centre," in *Proceedings of the 2003 Winter Simulation Conference*, Louisiana, 2003, pp. 1805-1812.

Pot, Bhulai, and Koole, "A Simple Staffing Method for Multiskill Call Centres," *Manufacturing & Service Operations Management*, vol. 10, no. 3, pp. 421–428, 2008.

Reynolds, P. (2003). Call center staffing: the complete, practical guide to workforce management. Nashville, TN: The Call Center School Press.

Robbins, T. and Medeiros, D., "Evaluating Arrival Rate Uncertainty in Call Centres," in *Proceedings of the 2006 Winter Simulation Conference*, California, 2006, pp. 2180-2187.

Robbins, T. and Harrison, T., "A Simulation Based Scheduling Model for Call Centres with Uncertain Arrival Rates," in *Proceedings of the 2008 Winter Simulation Conference*, Miami, 2008, pp. 2884-2890.

Robbins, T., Medeiros, D., and Harrison, T., "Does The Erlang C Model Fit in Real Call Centres?," in *Proceedings of the 2010 Winter Simulation Conference*, Maryland, 2010, pp. 2853-2864.

Robbins, T. R., Medeiros, D. J., & Dum, P. (2006). Evaluating Arrival Rate Uncertainty In Call Call Centres. Winter Simulation Conference.

Ronning, G., "Maximum Likelihood Estimation of Dirichlet Distributions," *Journal of Statistical Computation and Simulation*, vol. 32, no. 4, pp. 215-221, 1989.

Ross, S., *Introduction to Probability Models*, Ninth Edition ed.: Academic Press, pp. 302-307 and 316-320, 2007.

Roubos, A., Bhulai, S., and Koole, G., "Flexible Staffing for Call Centres with Non Stationary Arrival Rates," VU University Amsterdam, Department of Mathematics, Working Paper 2011.

Saltzman, R. and Mehrotra, V., "A Call Centre Uses Simulation to Drive Strategic Change," *Interfaces*, vol. 31, no. 3, pp. 87-101, 2001.

Sargent, R., "A New Statistical Procedure for Validation of Simulation and Stochastic Models," L.C. Smith College of Engineering and Computer Science, Technical Report 2010.

Sisselman, M. and Whitt, W., "Value-Based Routing and Preference-Based Routing in Customer Contact Centres," *Production and Operations Management*, vol. 16, no. 3, pp. 277–291, 2007.

Steckley, S., Henderson, S., and Mehrotra, V., "Performance Measures for Service Systems With a Random Arrival Rate," in *Proceedings of the 2005 Winter Simulation Conference*, Orlando, FL, 2005, pp. 566-575.

Tako, A. and Robinson, S., "Model Development in Discrete-Event Simulation and System Dynamics: An Empirical Study of Expert Modellers," *European Journal of Operational Research*, vol. 207, no. 2, pp. 784–794, 2010.

Taylor, P. & Huxley, S. (1989) A Break from Tradition for the San Francisco Police: Patrol Officer Scheduling Using an Optimization-Based Decision Support System. Interfaces,19 No.1,4-24

Tezcan, T. and Behzad, B., "Robust Design and Control of Call Centres with Flexible Interactive Voice Response Systems," *Manufacturing & Service Operations Management*, p. Published Online , May 2012.

Whitt, W., "Engineering Solution of a Basic Call-Centre Model," *Management Science*, vol. 51, no. 2, pp. 221–235, 2005.

Whitt, W., "Staffing a Call Centre with Uncertain Arrival Rate and Absenteeism," *Production and Operations Management*, vol. 15, no. 1, pp. 88–102, 2006.

Winston, W. L., and Venkataramanan, M. (2003) Introduction to Mathematical Programming.Operations Research: Volume 1.

Zohar, E., Mandelbaum, A., Shimkin, N., "Adaptive Behavior of Impatient Customers in Tele-Queues: Theory and Empirical Support," *Management Science*, vol. 48, no. 4, pp. 566-583, 2002.

APPENDICES

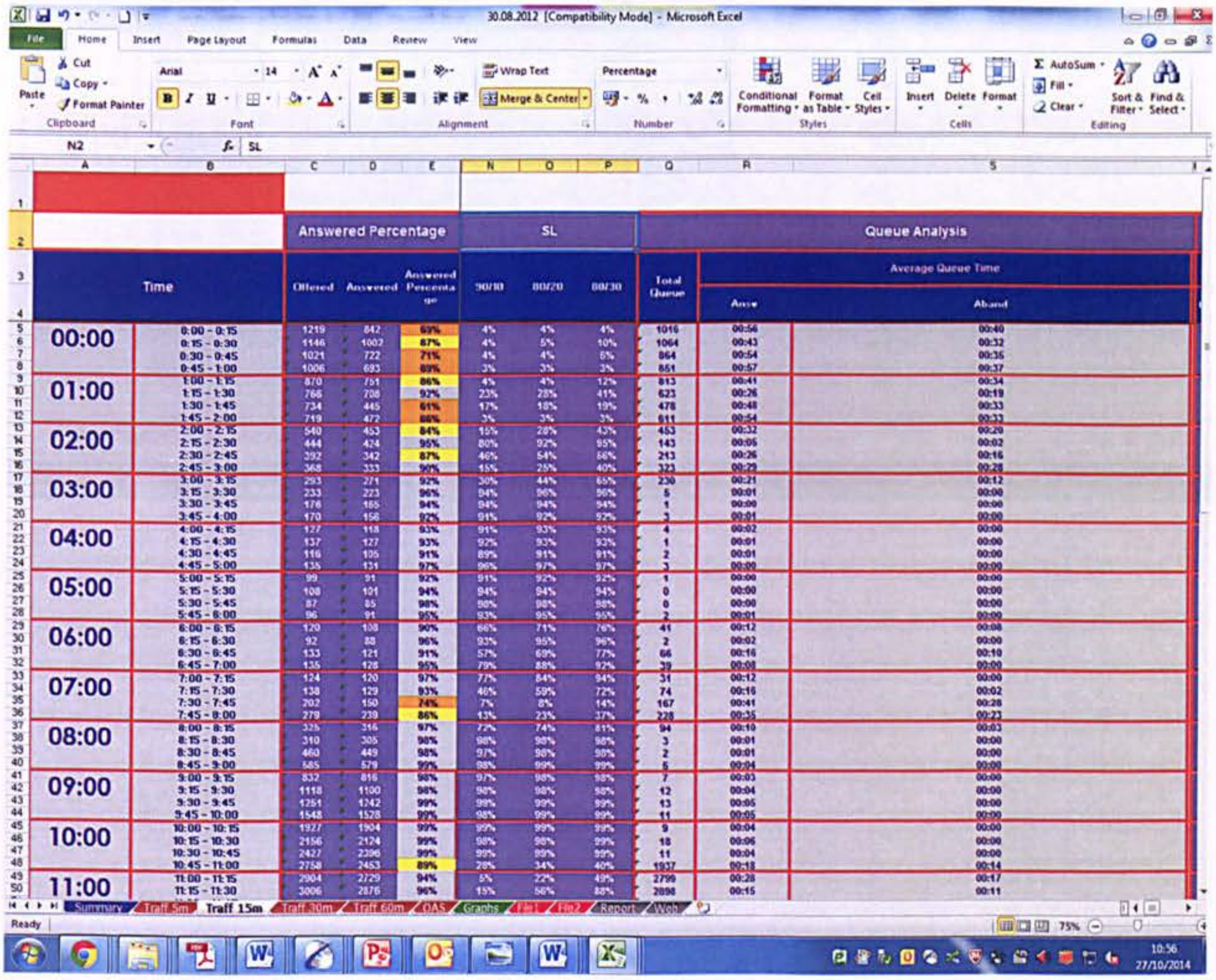
APPENDIX A: TYPICAL LINGO CODE

The following figure shows the LINGO code for solving the shift scheduling problem in the case of an-hour middle breaks with 2-hours time-window.

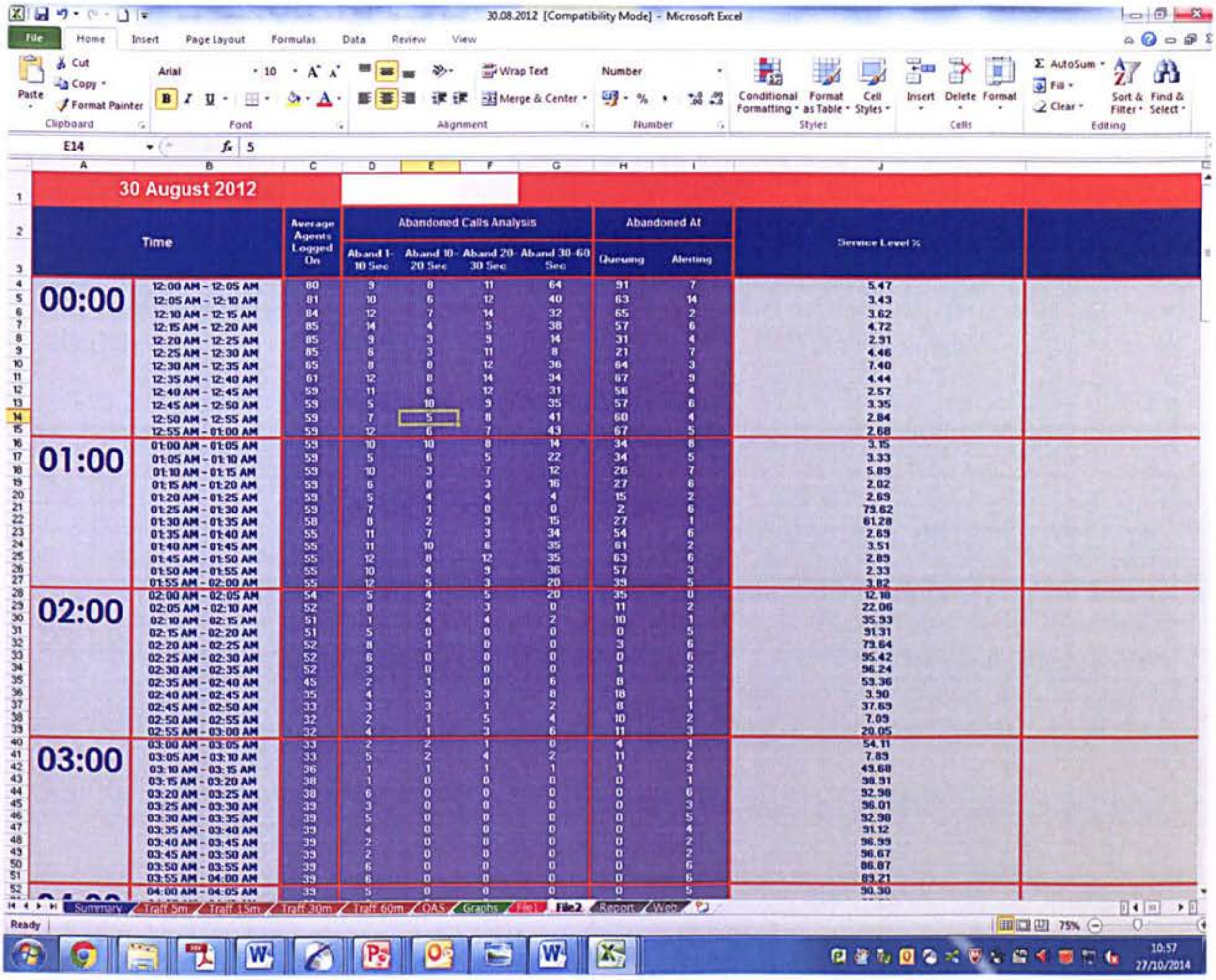
```
MODEL:
SETS:
Shift/1..j/:S;
Period/1..i/:P;
Assignment(Period,Shift):a;
Endsets
Data:
a =
1    1    0    0    0    0    0    0    0    0...................
1    1    0    0    0    0    0    0    0    0...................
1    1    1    1    0    0    0    0    0    0...................
1    1    1    1    0    0    0    0    0    0...................
1    1    1    1    1    1    0    0    0    0...................
1    1    1    1    1    1    0    0    0    0...................
1    1    1    1    1    1    1    1    0    0...................
1    1    1    1    1    1    1    1    0    0...................
1    1    1    1    1    1    1    1    1    1...................
1    1    1    1    1    1    1    1    1    1...................
1    1    1    1    1    1    1    1    1    1...................
1    1    1    1    1    1    1    1    1    1...................
1    0    1    1    1    1    1    1    1    1...................
1    0    1    1    1    1    1    1    1    1...................
1    0    1    0    1    1    1    1    1    1...................
1    0    1    0    1    1    1    1    1    1...................
0    1    1    0    1    0    1    1    1    1...................
0    1    1    0    1    0    1    1    1    1...................
0    1    0    1    1    0    1    0    1    1...................
0    1    0    1    1    0    1    0    1    1...................
1    1    0    1    0    1    1    0    1    0...................
1    1    0    1    0    1    1    0    1    0...................
1    1    1    1    0    1    0    1    1    0...................
1    1    1    1    0    1    0    1    1    0...................
1    1    1    1    1    1    0    1    0    1...................
1    1    1    1    1    1    0    1    0    1...................
1    1    1    1    1    1    1    1    0    1...................
1    1    1    1    1    1    1    1    0    1...................
1    1    1    1    1    1    1    1    1    1...................
1    1    1    1    1    1    1    1    1    1...................
1    1    1    1    1    1    1    1    1    1...................
1    1    1    1    1    1    1    1    1    1...................
.    .    .    .    .    .    .    .    .    ...................
.    .    .    .    .    .    .    .    .    ...................
Enddata

MIN = @SUM(Shift(j):S(j));
@FOR (Period(i):@SUM(Shift(j):a(i,j)*S(j))>=P(i));
End
```

APPENDIX B: SAMPLE OF EMPIRICAL DATA SHEETS

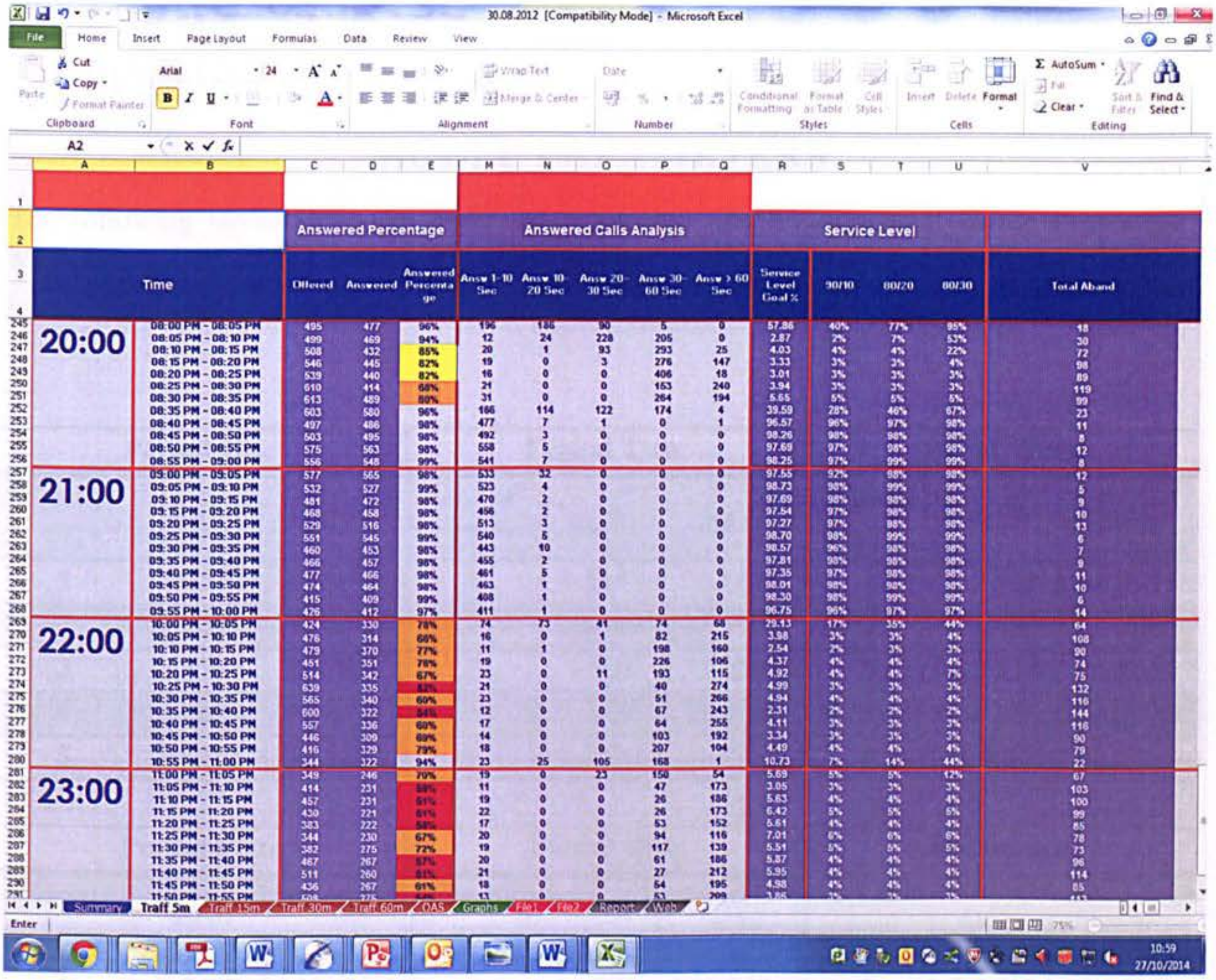
The following figure show a sample of obtained data sheets from the call center of the applied case study. Data sheet for August 30, 2012.



Time		Answered Percentage: Offered	Answered Percentage: Answered	Answered Percentage: Answered Percentage	SL: 90/10	SL: 80/20	SL: 80/30	Queue Analysis: Total Queue	Average Queue Time: Answ	Average Queue Time: Aband
00:00	0:00 - 0:15	1219	842	69%	4%	4%	4%	1016	00:56	00:40
	0:15 - 0:30	1146	1002	87%	4%	5%	10%	1064	00:43	00:32
	0:30 - 0:45	1021	722	71%	4%	4%	6%	864	00:54	00:35
	0:45 - 1:00	1006	693	69%	3%	3%	3%	851	00:57	00:37
01:00	1:00 - 1:15	870	751	86%	4%	4%	12%	813	00:41	00:34
	1:15 - 1:30	766	708	92%	23%	28%	41%	623	00:26	00:19
	1:30 - 1:45	734	445	61%	17%	18%	19%	478	00:48	00:33
	1:45 - 2:00	719	472	66%	3%	3%	3%	611	00:54	00:33
02:00	2:00 - 2:15	540	453	84%	15%	28%	43%	453	00:32	00:20
	2:15 - 2:30	444	424	95%	80%	92%	95%	143	00:05	00:02
	2:30 - 2:45	392	342	87%	46%	54%	66%	213	00:26	00:16
	2:45 - 3:00	368	333	90%	15%	25%	40%	323	00:29	00:28
03:00	3:00 - 3:15	293	271	92%	30%	44%	65%	230	00:21	00:12
	3:15 - 3:30	233	223	96%	94%	96%	96%	5	00:01	00:00
	3:30 - 3:45	176	165	94%	94%	94%	94%	1	00:00	00:00
	3:45 - 4:00	170	156	92%	91%	92%	92%	3	00:01	00:00
04:00	4:00 - 4:15	127	118	93%	91%	93%	93%	4	00:02	00:00
	4:15 - 4:30	137	127	93%	92%	93%	93%	1	00:01	00:00
	4:30 - 4:45	116	105	91%	89%	91%	91%	2	00:01	00:00
	4:45 - 5:00	135	131	97%	96%	97%	97%	3	00:00	00:00
05:00	5:00 - 5:15	99	91	92%	91%	92%	92%	1	00:00	00:00
	5:15 - 5:30	108	101	94%	94%	94%	94%	0	00:00	00:00
	5:30 - 5:45	87	85	98%	98%	98%	98%	0	00:00	00:00
	5:45 - 6:00	96	91	95%	93%	95%	95%	2	00:01	00:00
06:00	6:00 - 6:15	120	108	90%	66%	71%	76%	41	00:12	00:08
	6:15 - 6:30	92	88	96%	93%	95%	96%	2	00:02	00:00
	6:30 - 6:45	133	121	91%	57%	69%	77%	66	00:16	00:10
	6:45 - 7:00	135	128	95%	79%	88%	92%	39	00:08	00:00
07:00	7:00 - 7:15	124	120	97%	77%	84%	94%	31	00:12	00:00
	7:15 - 7:30	138	129	93%	46%	59%	72%	74	00:16	00:02
	7:30 - 7:45	202	150	74%	7%	8%	14%	167	00:41	00:28
	7:45 - 8:00	279	239	86%	13%	23%	37%	228	00:35	00:23
08:00	8:00 - 8:15	325	316	97%	72%	74%	81%	94	00:10	00:03
	8:15 - 8:30	310	305	98%	98%	98%	98%	3	00:01	00:00
	8:30 - 8:45	460	449	98%	97%	98%	98%	2	00:01	00:00
	8:45 - 9:00	585	579	99%	98%	99%	99%	6	00:04	00:00
09:00	9:00 - 9:15	832	816	98%	97%	98%	98%	7	00:03	00:00
	9:15 - 9:30	1118	1100	98%	98%	98%	98%	12	00:04	00:00
	9:30 - 9:45	1251	1242	99%	99%	99%	99%	13	00:05	00:00
	9:45 - 10:00	1548	1528	99%	98%	99%	99%	11	00:05	00:00
10:00	10:00 - 10:15	1927	1904	99%	99%	99%	99%	9	00:04	00:00
	10:15 - 10:30	2156	2124	99%	98%	98%	99%	18	00:06	00:00
	10:30 - 10:45	2427	2396	99%	99%	99%	99%	11	00:04	00:00
	10:45 - 11:00	2758	2453	89%	28%	34%	40%	1937	00:18	00:14
11:00	11:00 - 11:15	2904	2729	94%	5%	22%	49%	2799	00:28	00:17
	11:15 - 11:30	3006	2876	96%	15%	56%	88%	2898	00:15	00:11



Time		Average Agents Logged On	Abandoned Calls Analysis: Aband 1-10 Sec	Aband 10-20 Sec	Aband 20-30 Sec	Aband 30-60 Sec	Abandoned At: Queuing	Alerting	Service Level %
00:00	12:00 AM - 12:05 AM	80	9	8	11	64	91	7	5.47
	12:05 AM - 12:10 AM	81	10	6	12	40	63	14	3.43
	12:10 AM - 12:15 AM	84	12	7	14	32	65	2	3.62
	12:15 AM - 12:20 AM	85	14	4	5	38	57	6	4.72
	12:20 AM - 12:25 AM	85	9	3	9	14	31	4	2.91
	12:25 AM - 12:30 AM	85	6	3	11	8	21	7	4.46
	12:30 AM - 12:35 AM	65	8	8	12	36	64	3	7.40
	12:35 AM - 12:40 AM	61	12	8	14	34	67	9	4.44
	12:40 AM - 12:45 AM	59	11	6	12	31	56	4	2.57
	12:45 AM - 12:50 AM	59	5	10	9	35	57	6	3.35
	12:50 AM - 12:55 AM	59	7	5	8	41	60	4	2.84
	12:55 AM - 01:00 AM	59	12	6	7	43	67	5	2.68
01:00	01:00 AM - 01:05 AM	59	10	10	8	14	34	8	3.15
	01:05 AM - 01:10 AM	59	5	6	5	22	34	5	3.33
	01:10 AM - 01:15 AM	59	10	3	7	12	26	7	5.89
	01:15 AM - 01:20 AM	59	6	8	3	16	27	6	2.02
	01:20 AM - 01:25 AM	59	5	4	4	4	15	2	2.69
	01:25 AM - 01:30 AM	59	7	1	0	0	2	6	79.62
	01:30 AM - 01:35 AM	58	8	2	3	15	27	1	61.28
	01:35 AM - 01:40 AM	55	11	7	3	34	54	6	2.69
	01:40 AM - 01:45 AM	55	11	10	6	35	61	2	3.51
	01:45 AM - 01:50 AM	55	12	8	12	35	63	6	2.89
	01:50 AM - 01:55 AM	55	10	4	9	36	57	3	2.33
	01:55 AM - 02:00 AM	55	12	5	3	20	39	5	3.82
02:00	02:00 AM - 02:05 AM	54	5	4	5	20	35	0	12.18
	02:05 AM - 02:10 AM	52	8	2	3	0	11	2	22.06
	02:10 AM - 02:15 AM	51	1	4	4	2	10	1	35.93
	02:15 AM - 02:20 AM	51	5	0	0	0	0	5	91.31
	02:20 AM - 02:25 AM	52	8	1	0	0	3	6	79.64
	02:25 AM - 02:30 AM	52	6	0	0	0	0	6	95.42
	02:30 AM - 02:35 AM	52	3	0	0	0	1	2	96.24
	02:35 AM - 02:40 AM	45	2	1	0	6	8	1	53.36
	02:40 AM - 02:45 AM	35	4	3	3	8	18	1	3.90
	02:45 AM - 02:50 AM	33	3	3	1	2	8	1	37.69
	02:50 AM - 02:55 AM	32	2	1	5	4	10	2	7.09
	02:55 AM - 03:00 AM	32	4	1	3	6	11	3	20.05
03:00	03:00 AM - 03:05 AM	33	2	2	1	0	4	1	54.11
	03:05 AM - 03:10 AM	33	5	2	4	2	11	2	7.89
	03:10 AM - 03:15 AM	36	1	1	1	1	1	3	49.68
	03:15 AM - 03:20 AM	38	1	0	0	0	0	1	98.91
	03:20 AM - 03:25 AM	38	6	0	0	0	0	6	92.98
	03:25 AM - 03:30 AM	39	3	0	0	0	0	3	96.01
	03:30 AM - 03:35 AM	39	5	0	0	0	0	5	92.98
	03:35 AM - 03:40 AM	39	4	0	0	0	0	4	91.12
	03:40 AM - 03:45 AM	39	2	0	0	0	0	2	96.99
	03:45 AM - 03:50 AM	39	2	0	0	0	0	2	96.67
	03:50 AM - 03:55 AM	39	6	0	0	0	0	6	86.87
	03:55 AM - 04:00 AM	39	6	0	0	0	0	6	89.21
	04:00 AM - 04:05 AM	39	5	0	0	0	0	5	90.30


30.08.2012 [Compatibility Mode] - Microsoft Excel
Answered Percentage
Answered Calls Analysis
Service Level
Time
Offered
Answered
Answered Percentage
Answ 1-10 Sec
Answ 10-20 Sec
Answ 20-30 Sec
Answ 30-60 Sec
Answ > 60 Sec
Service Level Goal %
90/10
80/20
80/30
Total Aband
20:00
21:00
22:00
23:00
Summary
Traff 5m
Traff 15m
Traff 30m
Traff 60m
QAS
Graphs
File1
File2
Report
Web

APPENDIX C: ABANDONMENT GROUP PROPORTIONS MODEL PARAMETERS

The following tables show the estimated parameters for Abandonment Group Proportion Model.

Period One			Period Two			Period Three		
Group (1)	$\alpha_b(1)$	$\beta_b(1)$	Group (2)	$\alpha_b(2)$	$\beta_b(2)$	Group (3)	$\alpha_b(3)$	$\beta_b(3)$
1	1.6579	97.9974	1	1.5148	73.1127	1	1.5504	66.0175
2	1.3691	98.2863	2	1.0509	73.5766	2	0.7384	66.8295
3	2.1727	97.4826	3	1.568	73.0595	3	1.0814	66.4864
4	2.3429	97.3125	4	1.5536	73.0739	4	0.969	66.5989
5	9.1258	90.5295	5	2.8503	71.7772	5	1.4262	66.1417

Period Four			Period Five			Period Six		
Group (4)	$\alpha_b(4)$	$\beta_b(4)$	Group (5)	$\alpha_b(5)$	$\beta_b(5)$	Group (6)	$\alpha_b(6)$	$\beta_b(6)$
1	4.8246	149.5093	1	5.5585	137.6747	1	8.6203	210.3139
2	0.3431	153.9908	2	0.2981	142.9351	2	0.3126	218.6216
3	0.3233	154.0106	3	0.2135	143.0198	3	0.1831	218.7511
4	0.2516	154.0823	4	0.1879	143.0453	4	0.1724	218.7618
5	0.2458	154.0881	5	0.1879	143.0453	5	0.1724	218.7618

Period Seven			Period Eight			Period Nine		
Group (7)	$\alpha_b(7)$	$\beta_b(7)$	Group (8)	$\alpha_b(8)$	$\beta_b(2)$	Group (9)	$\alpha_b(9)$	$\beta_b(3)$
1	4.7199	148.1043	1	2.4687	100.7746	1	10.4732	695.0490
2	0.4296	152.3946	2	0.4677	102.7757	2	0.6161	704.9061
3	0.2201	152.6041	3	0.2392	103.0041	3	0.3356	705.1866
4	0.2144	152.6098	4	0.2355	103.0078	4	0.2136	705.3086
5	0.2151	152.6091	5	0.3157	102.9277	5	0.3148	705.2074

Period Ten			Period Eleven			Period Twelve		
Group (10)	$\alpha_b(10)$	$\beta_b(10)$	Group (11)	$\alpha_b(11)$	$\beta_b(11)$	Group (12)	$\alpha_b(12)$	$\beta_b(12)$
1	20.4134	1885.6091	1	1.7909	156.0258	1	1.4639	144.0273
2	0.7671	1905.2553	2	0.8669	156.9497	2	2.1488	143.3424
3	0.3959	1905.6265	3	0.732	157.0847	3	1.7421	143.7491
4	0.3301	1905.6923	4	0.4084	157.4082	4	1.4229	144.0683
5	0.4166	1905.6058	5	0.4741	157.3426	5	2.4515	143.0396

Period Thirteen			Period Fourteen			Period Fifteen		
Group (13)	$\alpha_b(13)$	$\beta_b(13)$	Group (14)	$\alpha_b(14)$	$\beta_b(14)$	Group (15)	$\alpha_b(15)$	$\beta_b(15)$
1	0.8807	67.3437	1	0.743	42.6640	1	0.6879	33.2169
2	1.1176	67.1068	2	0.8694	42.5376	2	0.509	33.3958
3	1.2137	67.0106	3	0.7073	42.6997	3	0.3783	33.5266
4	1.1277	67.0967	4	0.6317	42.7753	4	0.3142	33.5907
5	2.1584	66.0660	5	0.8028	42.6042	5	0.3533	33.5515

Period Sixteen		
Group (16)	$\alpha_b(16)$	$\beta_b(16)$
1	0.9835	58.6135
2	0.6994	58.8975
3	0.5805	59.0165
4	0.4919	59.1051
5	0.5263	59.0706

Period Seventeen		
Group (17)	$\alpha_b(17)$	$\beta_b(17)$
1	1.8444	129.6887
2	0.9411	130.5920
3	0.6144	130.9187
4	0.5277	131.0054
5	0.6186	130.9145

Period Eighteen		
Group (18)	$\alpha_b(18)$	$\beta_b(18)$
1	4.2594	289.9351
2	0.4347	293.7598
3	0.2494	293.9451
4	0.2732	293.9214
5	0.3216	293.8729

Period Nineteen		
Group (19)	$\alpha_b(19)$	$\beta_b(19)$
1	6.8358	405.5950
2	0.5509	411.8800
3	0.2755	412.1553
4	0.2349	412.1960
5	0.2787	412.1521

Period Twenty		
Group (20)	$\alpha_b(20)$	$\beta_b(20)$
1	7.9204	535.8213
2	0.956	542.7857
3	0.3827	543.3590
4	0.2785	543.4632
5	0.2674	543.4743

Period Twenty-One		
Group (21)	$\alpha_b(21)$	$\beta_b(21)$
1	1.1369	58.4713
2	0.6021	59.0061
3	0.4391	59.1691
4	0.2821	59.3262
5	0.2975	59.3108

Period Twenty-Two		
Group (22)	$\alpha_b(22)$	$\beta_b(22)$
1	1.604	91.8362
2	0.5992	92.8410
3	0.3879	93.0523
4	0.3348	93.1055
5	0.2734	93.1668

Period Twenty-Three		
Group (23)	$\alpha_b(23)$	$\beta_b(23)$
1	0.7698	29.8058)
2	0.4726	30.1030)
3	0.3858	30.1898)
4	0.2906	30.2850)
5	0.3847	30.1909)

Period Twenty-Four		
Group (24)	$\alpha_b(24)$	$\beta_b(24)$
1	0.5799	18.0064)
2	0.4426	18.1437)
3	0.3175	18.2689)
4	0.2955	18.2908)
5	0.4107	18.1756)