Maximizing Patient Satisfaction in Systems with Time-Varying Arrival Rates

A thesis submitted to the
College of Graduate and Postdoctoral Studies
in partial pulfillment of the requirements
for the degree of Master of Science
in the Department of Mathematics and Statistics
University of Saskatchewan
Saskatoon

By
Leila Rabiei Fard

Permission to Use

Head of the Department of Mathematics & Statistics
142 McLean Hall
106 Wiggins Road
University of Saskatchewan
Saskatoon, Saskatchewan
Canada
S7N 5E6

OR

Dean
College of Graduate and Postdoctoral Studies
University of Saskatchewan

116 Thorvaldson Building, 110 Science Place
Saskatoon, Saskatchewan S7N 5C9 Canada

Abstract

Time-Varying Little's Law (TVLL) can be regarded as part of the theory of Infinite Servers (IS) models, for the abstract system can be considered as a general IS model if waiting time is considered as service time. Moreover, the time-varying arrival rate does not affect the waiting time distribution, when there are adequate time-varying servers in the system. In this study, we estimate the average number of entities in the system over a sub-interval and the arrival rate function, and apply TVLL combined with time-varying staffing to estimate the unknown mean wait times. When the arrival rate function is approximated by a linear (quadratic) function, the average waiting time satisfies a quadratic (cubic) equation. The estimation of average waiting time based on TVLL is a positive real root of the average waiting time equation.

If, the arrival rate function is neither approximately linear nor approximately quadratic, it must be approximated by a polynomial function of higher degree. In this study, we investigate systems with arrival rate function of degree 3, and find the estimation of average waiting time which is the root of a polynomial of degree 4.

Also, we study queues with time-varying arrival rate to obtain optimal visit time leading to maximum satisfaction of patients in walk-in clinics. If there is adequate time-varying staffing, then customers receive service upon arrival and waiting times tend to be approximately as equal as the service times though the arrival rates are time-varying. However, in the systems with limited servers, some customers must wait in the waiting room and when there is no room in the area, the new arriving customers are refused. Rejection of customers may lead to their dissatisfaction. If we decrease the average service time, less customers will be refused, but shorter service time decreases happiness of admitted customers.

Another issue is the revenue of walk-in clinics. Walk-in clinics work on a fee-for-service model, so they benefit from the number of patients they serve. As the number of patients increases, more revenue is gained. Hence, it may be in interest of some walk-in clinics to reduce visit times to increase profit. As mentioned, short visit time sacrifices the quality of service and leads to the dissatisfaction of patients. Patients want to be heard carefully and be asked directly why they have come to the clinic. The problem gets worse in rush hours when the number of arrivals increases but the number of servers could not be increased due to limitation in the number of doctors.

We obtain optimal value for visit time considering satisfaction of customers and revenue of walk-in clinics simultaneously.

Keywords: Queueing System, Markov Chain, Queueing Theory, Little's law (LL), Time-Varying Little's Law (TVLL), Time-Varying Arrival Rate

Acknowledgements

I am extremely grateful to my supervisors, Dr. Ebrahim Samei and Dr. Hamed Samarghandi for their invaluable advice, continuous support, and patience during my study. I deeply appreciate how they have been continuously encouraging and guiding me in the last two years, and also how they have always been so friendly and supportive of all of my efforts and struggles. Working under their supervision has been very enjoyable and I have learned and grown a lot.

Contents

List of Tables

List of Figures

List of Abbreviations

CM	Capacitated Monopolistic
CO	Capacitated Oligopoly
EDA	Exploratory Data Analysis
ED	Emergency Department
EW	Emergency Ward
FCFS	First Come First Served
FIFO	First In First Out
HOL	Head of Line
IS	Infinite Servers
IWLS	Iterative Weighted Least Square
KS	Kolmogorov–Smirnov
LCFS	Last Come First Served
LIFO	Last Come First Out
LL	Little's Law
MCE	Mass Casualty Event
ML	Maximum Likelihood
NHPP	Non-Homogeneous Poisson Process
NSPP	Non-Stationary Poisson arrival Process
NSNP	Non-Stationary Non-Poisson
OLS	Ordinary Least Square
OR	Operations Research
PS	Processor Sharing
QoS	Quality of Service
SIRO	Service In Random Order
RS	Random Service
TVLL	Time Varying Little's Law
UM	Uncapacitated Monopolistic

UO	Uncapacitated Oligopoly

1. Introduction

The vast majority of people have had the unpleasant experience of waiting in the long queues in their lifetime. Not only customers but also business owners are influenced by long waiting times, because long queues lead to undesirable experience which damages clients' loyalty ([1], [2]). This shows the importance of studying models and techniques to analyse queues and reduce the duration of waiting times. Queuing Theory, which is a tool for analysis of technical problems, deals with waiting time. The applications of Queuing Theory are apparent in many fields such as manufacturing systems, computer and communication networks, transportation networks, service management, supply chain management, sharing economics, healthcare and so forth. Evidence shows that the results obtained from studying queuing theory improve efficiency and profit significantly. Hence, studying queuing is very important and practical.

1.1 Motivation

In comparison with research on the behavior of queues with constant rates, far less literature exists on time dependent types of queues [3]. This is due to greater mathematical complexity of time-varying rate problems. Moreover, many of the theoretical tools such as equilibrium probabilities for Markov chains, matrix geometric solutions, and Laplace transforms are not available or directly applicable for queues with time-varying rates. The results obtained from working in this field can help to determine which mathematical tools are required to advance the theory of non-stationary queues. Creating such a new theory provides new formulas and algorithms to employ in the performance modelling of queuing systems especially in healthcare centers. In addition, healthcare centers including walk-in clinics are complex systems due to time dependent behavior of their queuing models. Studying queues with non-stationary parameters is very useful for enhancing the performance of walk-in clinics which have an incredible role in improving the health of the community and have essential societal and economical benefits.

1.2 Research framework

The main purpose in this research is to obtain optimum value for visit time in walk-in clinics with time-varying arrival rate. First, we will discuss the different approximations for time-varying arrival rates and waiting times. Then, according to the unique situation of each walk-in clinic in the area under study, we define different scenarios. In these scenarios, we consider time-varying arrival rates and limited number of servers and obtain optimum visit time with considering patients' satisfaction and clinics' revenue.

1.2.1 Time-varying arrival rate

A queueing system is a system with a service facility at which customers arrive for service. The arrival rate is simply how many arrivals occur in a specified length of time interval. In reality, it cannot be considered constant everywhere. For example in taxi stations, banks, call and healthcare centers, the number of arrivals changes at different time intervals. The time dependent function of arrival rate $\lambda(s)$ could be approximated with different functions. In this research, we suppose three polynomial functions to estimate arrival rates:

$$\lambda(s) \approx \lambda_l(s) = a + bs \qquad 0 \leq s \leq t$$

$$\lambda(s) \approx \lambda_q(s) = a + bs + cs^2 \qquad 0 \leq s \leq t$$

$$\lambda(s) \approx \lambda_c(s) = a + bs + cs^2 + ds^3 \qquad 0 \leq s \leq t.$$

In the next step, we obtain waiting time in an ideal system with time-varying staff where the number of severs changes as the arrival rate changes. In addition, the performance of an ideal system is similar to IS system (Infinite Servers system). In the real world, however, it is not always possible to provide enough servers such that all arrivals receive servers upon arrival. In this research, queuing systems with limited number of service providers will be investigated.

1.2.2 Service time in walk-in clinics

Some walk-in clinics accept all patients because of their unlimited capacity, while others have to reject some patients due to their capacitated waiting room. Apart from this, some clinics act in a monopoly condition, whereas others serve the patients in an oligopoly condition. So, we should separate clinics based on their individual capacity and and market structures. We define four scenarios:

- Scenario 1 (Model UM): an Uncapacitated walk-in clinic in a Monopolistic market, where there are no other walk-in clinics in the area. The model is named Model UM, where U and M show the Uncapacitated capacity and Monopolistic position of the clinic in the region, respectively.

- Scenario 2 (Model CM): a Capacitated walk-in clinic where a certain number of patients can be served. In this scenario, the clinic operates in a Monopolistic market.

- Scenario 3 (Model UO): This scenario considers an Uncapacitated walk-in clinic in an Oligopoly market. In other words, there are several competitors in the vicinity.

- Scenario 4 (Model CO): In this model, a Capacitated walk-in clinic is considered in an area in which there are some other walk-in clinics. The model is called CO to describe the Uncapacitated capacity and Oligopolistic position of the clinic in the region.

Then, optimum visit time is obtained based on satisfaction of patients and revenue in each scenario.

1.3 Thesis organization

This dissertation comprises seven chapters as listed below:

- Chapter 1 investigates the reasons motivating our work. Also, it contains the research framework to provide a clear idea what exactly the problem is and what is done about it.

- Chapter 2 is a concise overview of research that has been done on queues with time-varying arrival rates, largely focusing on healthcare centers. This brief overview shows the gaps and areas needing further study.

- Chapter 3 provides fundamentals of queuing modelling and gives an introduction to non-homogeneous Poisson processes (NHPPs) which are widely used to model time-dependent queues in a multitude of stochastic models. In this chapter, statistical analysis with time-varying Little's Law (TVLL) will be reviewed to estimate waiting time in a time-varying staffing system. Also, linear and quadratic approximation of arrival rates will be described and a new approximation of arrival rates in the form of cubic polynomial function will be introduced.

- Chapter 4 introduces a new version of Model UM with a limited number of servers and investigates how to find the optimum value of service time maximizing satisfaction of patients. In this chapter, Model CM is explained in detail for multi-server queues and a new revenue function ? is introduced which considers patients' satisfaction and clinics' profit simultaneously.

- Chapter 5 studies Model UO and Model CO. Satisfaction of patients and revenue in an oligopoly situation will be discussed.

- Chapter 6 discusses the implementation of the different Models including UM, CM, UO, and CO. Since the closed-form solution of the patient satisfaction maximization problem could not be found, numerical example would be defined. Also, sensitivity to different parameters will be analyzed.

- In chapter 7, we summarize the main results of the research and list potential future research directions.

2. Literature review

This chapter provides a thorough review of the researchers' work on queues with time-varying arrival rates and service time in healthcare organizations. Since different concepts and subjects will be reviewed, the literature review chapter is divided into some sub-sections, each dedicated to a specific subject.

2.1 Queues with time-varying arrival

The study of queues with time-varying rates started in mid-twentieth century. A remarkable early work has been done by E Brockmeyer et al. [4]. Two other early works have been done by Rothkopf and Oren [5] and Newell [6] that inspired the Messey's Ph.D. thesis [7] in non-stationary queues, under the direction of Joseph B. Keller.

In recent years, studying these types of queues has increased significantly because parameters are non-stationary in real world and for modeling queues in reality, we need to understand the behavior of non-stationary queues. Most research on time-varying arrival queues has focused on arrival process models and time-varying Little's Law (TVLL). In this section, we will review literature on arrival process models and TVLL.

2.1.1 Arrival process models

When building stochastic models for enhancing the performance of service systems, it is important to have an appropriate arrival process model. For modeling queuing systems, usually stationary (homogeneous) Poisson arrivals are assumed. However, development of new technologies and appearance of new services have led to new traffic models with non-stationary parameters. In addition, the classical Poisson process traffic model could not be applied when the arrival rate changes considerably over time intervals [8].

The natural arrival process model when there is a time-varying arrival rate is a non-homogeneous Poisson process (NHPP). Ways to test and model non-Poisson and non-stationary arrival processes have been studied in Massey and Whitt [9], Gebhardt and Nelson [10], Nelson and Gerhardt [11], and Zhang et al [12]. In addition, Massey and Whitt [9] made a connection between laws of large numbers and central limit theorems for non-stationary counting processes to corresponding limits for their inverse processes. Then, these results were applied to develop approximations for queues that are unstable in a non-stationary manner. For modeling, they constructed non-stationary point processes as random time-transformations of familiar point processes, like renewal processes and stationary point processes. Gebhardt and Nelson [10] extended techniques that transform a stationary Poisson arrival process into a non-stationary Poisson arrival process (NSPP) by transforming a stationary renewal process into a non-stationary, non-Poisson (NSNP) arrival process. They illustrated that the desired arrival rate is obtained and that when the renewal base process is either more or less variable than Poisson, then the NSNP process is also more or less variable, respectively, than an NSPP. They also suggested methods for specifying the renewal base process when presented properties of, or data from, an arrival process and showed them by modeling real arrival data. In 2011, Nelson and Gerhardt [11] introduced another technique to model and simulate non-stationary, non-renewal arrival processes depending merely on the analyst setting intuitive and easily controllable parameters which was suitable for assessing the effect of non-stationary, non-exponential, and non-independent arrivals on simulated performance when they are suspected. Zhang et al. [12] identified a significant factor characterizing the stochastic variability of the arrivals to their averages which was referred as the scaling parameter having a profound impact on the design of staffing rules. To capture the scaling parameter a new model was proposed.

Ways to fit or approximate the arrival rate function were studied in Massey et al. [13] and Massey and Whitt [14]. Furthermore, Massey et al. [13] estimated the parameters of a non-homogeneous Poisson process with linear rate over a finite interval. Investigated ways were ordinary least squares (OLS), iterative weighted least squares (IWLS) and maximum likelihood (ML). Also, statistical tests to determine whether the linear Poisson model is appropriate were developed.

Massey and Whitt [14] considered $M_t/G/s/0$ model with a non-homogeneous Poisson arrival process. They proposed a specific approximation based on the heavy-traffic peakedness formula.

2.1.2 Little's Law and time-varying version

In a paper published in 1954 [15], Little's Law was assumed true and used without proof [16]. In 1961, John D. C. Little [17] published a paper in Operations Research to proof the law. Over the years, this formula has become widely known as Little's Law. Due to its theoretical and practical importance, this formula is now very well known in Queuing Theory.

Whitt [18] stated that there is greater unity in the overall theory than had been previously realized. He emphasized the fundamental Little's Law is intimately connected to the infinite server (IS) queuing model. In addition, the IS model with a time-varying arrival rate is in turn connected to the time-varying Little's law (TVLL) as discussed in Bertsimas and Mourtzinou [19], Fralix and Riano [20], and Kim and Whitt [21]. Bertsimas and Mourtzinou [19] established a transient Little's law at the same level of generality as the classical stationary version of Little's law. Then, they obtained transient distributional laws for overtaking free non-stationary systems. These laws relate the distributions of the number of customers in the system and the delay at time t and constitute a complete set of equations that describes the dynamics of overtake free non-stationary queuing systems. Moreover, they extended these laws to multi-class systems as well. Finally, to demonstrate the power of the transient laws, they applied them to a variety of queuing systems: Infinite and single server systems with non-stationary Poisson arrivals and general non-stationary services, multi-class single server systems with general non-stationary arrivals and services, and multi-server systems with renewal arrivals and deterministic services, operating in the transient domain. For all specific systems they related the performance measures using the established set of laws and obtained a complete description of the system in the sense that they have a sufficient number of integral equations and unknowns. They then solved the set of integral equations using asymptotic expansions and exact numerical techniques. Finally, they reported computational results from their suggested methods. Fralix and Riano [20] took a new look at transient or time-dependent Little laws for queueing systems. Through the use of Palm measures, they represented that previous laws (see Bertsimas and Mourtzinou [19]) can be generalized. Furthermore, within this framework, a new law can be derived as well, which gives higher-moment expressions for very general types of queueing system; in particular, the laws hold for systems that allow customers to overtake one another. What is especially novel about their approach is the use of Palm measures that are induced by non-stationary point processes, as these measures are not commonly found in the queuing literature. This new higher-moment law is then used to provide expressions for all moments of the number of customers in the system in an $M/G/1$ preemptive last-come-first-served queue at a time $t > 0$, for any initial condition and any of the more famous preemptive disciplines (i.e. preemptive-resume, and preemptive-repeat with and without resampling) that are analogous to the special cases found in Abate and Whitt [22], [23]. These expressions are then used to derive a nice structural form for all of the time-dependent moments of a regulated Brownian motion (see Abate and Whitt [24], [25]). Kim and Whitt [21] stated that TVLL can be regarded as part of the theory for IS models, because the abstract system can be regarded as a general IS model, if we simply call the waiting time as the service time in the IS model. They concentrated on the application of TVLL to estimate waiting times by fitting a linear and quadratic function to arrival data. They also showed that the bias in the simple indirect estimator can be estimated and reduced by applying the time-varying Little's law (TVLL).

Little's law can be important for estimation, as shown in Glynn and Whitt [26], Lovejoy and Desmond [27], Kim and Whitt [28]. Moreover, for a large class of queuing systems, Little's law provides a variety of statistical estimators for the long-run time-average queue length and the long-run customer-average waiting time. Glynn and Whitt [26] applied central limit theorem versions of Little's law to investigate the asymptotic efficiency of these estimators. It was shown that an indirect estimator for time-average queue length using the natural estimator for waiting time plus the known arrival rate is more efficient than a direct estimator for time-average queue length, provided that the inter-arrival and waiting times are negatively correlated. They also introduced a general framework for indirect estimation which can be applied to other problems besides fundamental little's law. They showed that the issue of indirect-versus-direct estimation is related to estimation using nonlinear control variables and under mild regularity conditions, that any nonlinear control-variable scheme is equivalent to a linear control-variable scheme from the point of view of asymptotic efficiency. In addition, they indicated that asymptotic bias is typically asymptotically negligible compared to asymptotic efficiency. Lovejoy and Desmond [27] used Little's Law for estimating appropriate size for the observation unit in health care organizations and a natural internal consistency check on data. Furthermore, they stated expanding hospital capacity by developing an observation would be a significant strategy in congested hospitals. Understanding the principles for evaluating the potential impact and appropriate sizing of an observation unit is important. Lovejoy and Desmond [27] contrasted two approaches to determining observation unit sizing and profitability, real options, and a flow analysis based on Little's Law.

Applications of fundamental Little's Law with actual system data involve measurements over a finite-time interval. Kim and Whitt [28] investigated how estimates of number of customers in the system and arrival rate can be used to estimate waiting time when the waiting times are not observed. They advocated estimating confidence intervals. Given a single sample-path segment, they suggest estimating confidence intervals using the method of batch means, as is often done in stochastic simulation output analysis. Finally, they indicated how to estimate and remove bias due to interval edge effects when the system does not begin and end empty.

A sample-path version of a periodic Little's law has recently been established in Whitt and Zhang [29], which is motivated by the data analysis of an Israeli emergency department in Whitt and Zhang [30].

2.2 Healthcare

High-quality health plays a key role in human happiness and well-being contributing considerably to prosperity, wealth and even economic progress, as healthy populations are more productive, save more and live longer. This shows the importance of studying healthcare centers which are complex systems with essential societal benefits and huge mounting costs. Decision-makers should concentrate on improving healthcare quality even in ever-increasing pressures to make sure all people get the healthcare services they need. On the other hand, if they concentrate on improving healthcare, they will face challenges such as rising costs, lower reimbursements, and new regulatory demands. By studying and modeling processes in healthcare centers, we can deliver great insight for decision-makers to make the best decision considering quality of service and revenue simultaneously.

2.2.1 Healthcare organizations and time-varying arrival rate

Typically, healthcare centers have strongly time-varying arrivals and generally a non-homogeneous Poisson processes (NHPP) model is considered for such arrival process. However, this model should be tested by applying appropriate statistical tests to arrival data. Assuming that the NHPP has a rate that can be regarded as approximately piecewise-constant, a Kolmogorov–Smirnov (KS) statistical test of a Poisson process (PP) can be applied to test for a NHPP by combining data from separate sub-intervals, exploiting the classical conditional-uniform property. Kim and Whitt [31] applied KS tests to hospital emergency department arrival data and showed that they are consistent with the NHPP property, but only if that data is analyzed carefully. Initial testing rejected the NHPP null hypothesis because it failed to account for three common features of arrival data: (i) data rounding, (ii) choosing sub-intervals over which the rate varies too much, and (iii) over dispersion caused by combining data from fixed hours on a fixed day of the week over multiple weeks that do not have the same arrival rate. Kim and Whitt [31] investigated how to address each of these three problems.

Kim et al. [32] also applied statistical tests to arrival data from an endocrinology clinic, where arrivals are by appointment. The clinic data were also consistent with an NHPP within each day, but exhibit under-dispersion over multiple days. Kim et al. [32] introduced a new Gaussian-uniform arrival process model, with Gaussian daily totals and uniformly distributed arrivals given the totals.

In 2018, Kim et al. [33] developed a high-fidelity simulation model of the patient arrival process to an endocrinology clinic by carefully examining appointment and arrival data from that clinic. The used data included the time that the appointment was originally made as well as the time that the patient actually arrived, as well as if the patient did not arrive at all, in addition to the scheduled appointment time. They take a data-based approach, specifying the schedule for each day by its value at the end of the previous day. This data-based approach shows that the schedule for a given day evolves randomly over time. Indeed, in addition to three recognized sources of variability—(i) no-shows, (ii) extra unscheduled arrivals, and (iii) deviations in the actual arrival times from the scheduled times—they found that the primary source of variability in the arrival process is variability in the daily schedule itself. Even though service systems with arrivals by appointment can differ in many ways, their data-based approach to modeling the clinic arrival process was a guideline or template for constructing high-fidelity simulation models for other arrival processes generated by appointments.

Yom-Tov and Mandelbaum [34] analyzed a queueing model that is named Erlang-R, where the “R” stands for reentrant customers. Erlang-R accommodates customers who return to service several times during their sojourn within the system, and its modeling power is most pronounced in time-varying environments. Indeed, it was motivated by healthcare systems, in which offered-loads vary over time and patients often go through a repetitive service process. Erlang-R helps answer questions such as how many servers (physicians/nurses) are required to achieve predetermined service levels. Formally, it is merely a two-station open queuing network, which, in a steady state, evolves like an Erlang-C $M/M/k$ model. In time-varying environments, on the other hand, the situation differs: in these systems, one must account for the reentrant nature of service to avoid excessive staffing costs or undesirable service levels. Yom-Tov and Mandelbaum [34] validated Erlang-R against an emergency ward (EW) operating under normal conditions as well as during a mass casualty event (MCE). In both scenarios, they applied time-varying fluid and diffusion approximations: the EW is critically loaded and the MCE is overloaded. In particular, for the EW, a time-varying square-root staffing policy was proposed, based on the modified offered-load, which is proved to perform well over small-to-large systems.

A queuing-network view of patient flow in healthcare centers with non-stationary parameters is also very important for improving the performance of healthcare organizations. Armony et al. [35] explored patient flow data through the lens of a queuing scientist. The means is exploratory data analysis (EDA) in a large Israeli hospital, which reveals important features that are not readily explainable by existing models. Jacobson et al. [36] worked on allocation of scarce healthcare resources to improve patient flow, while minimizing health care delivery costs and increasing patient satisfaction. They suggested discrete-event simulation which is a popular and effective decision-making tool for optimal allocation. Moreover, combined optimization and simulation tools allow decision-makers to quickly determine optimal system configurations, even for complex integrated facilities. They provide an overview of discrete-event simulation modeling applications to health care clinics and integrated health care systems (e.g. hospitals, outpatient clinics, emergency departments, and pharmacies) over the past forty years.

Shi et al. [37] studied operations in the inpatient wards and their interface with the ED. Their main focus was on understanding the effect of inpatient discharge policies and other operational policies on the time-of-day waiting time performance, such as the fraction of patients waiting longer than six hours in the ED before being admitted. They proposed a novel stochastic processing network with the following characteristics to model inpatient operations:

- A patient's service time in the inpatient wards depends on that patient's admission and discharge times and length of stay. The service times capture a two-time-scale phenomenon and are not independent and identically distributed.

- Pre- and post-allocation delays model the extra amount of waiting caused by secondary bottlenecks other than bed unavailability, such as nurse shortage

- Patients waiting for a bed can overflow to a non primary ward when the waiting time reaches a threshold, where the threshold is time dependent

They showed, via simulation studies, that their model is able to capture the inpatient flow dynamics at hourly resolution and can evaluate the impact of operational policies on both the daily and time-of-day waiting time performance. In particular, their model predicts that implementing a hypothetical policy can eliminate excessive waiting for those patients who request beds in mornings. This policy incorporated the following components: a discharge distribution with the first discharge peak between 8 a.m. and 9 a.m. and 26% of patients discharging before noon, and constant-mean allocation delays throughout the day. The insights gained from their model can help hospital managers to choose among different policies to implement depending on the choice of objective, such as to reduce the peak waiting in the morning or to reduce daily waiting time statistics.

2.2.2 Service time in healthcare centers

Visit time plays a key role in satisfaction of patients. This fact has been investigated by two groups of researchers.

First group , including Fenton et al. [38], Schwartz et al. [39], and Gross et al. [40], considered visit time directly in their surveys. In addition, Fenton et al. [38] assumed visit time as one of the four factors pertaining to physician communication. Schwartz et al. [39] stated a high score to the time, service providers (doctors and nurses) allocate to them to hear their problem. Gross et al. [40] got a result that visit time is a significant factor in patient satisfaction.

Second group, Boudreaux and O'Hea [41], Boudreaux et al . [42], Jackson at el. [43], assumed visit time indirectly in their surveys. They recognized mutual communication of staff, specially the doctors, as one of the most critical factors to evaluate patient satisfaction. Boudreaux and O'Hea [41] cited interpersonal interactions with the doctor as the most important factor in patient satisfaction. Interpersonal communication represents physician's manner and the amount, quality and understand-ability of information transferred by the doctor to the patient. In another survey, Boudreaux et al. [42] set a list of significant factors for patient satisfaction including

- how well the doctor explains the plan of care,
- shows interest in patients concerns,
- conveys information, and
- guides the patient for homecare.

Jackson et al. [43] named the patient-doctor communication, including the information conveyed, as one of the factors in post-visit patient satisfaction. In spite of the fact that this group of researches do not directly consider visit time as a patient satisfaction factor, they consider time-consuming actions such as amount of information conveyed to the patients as important drivers of patient satisfaction. Hence, it can be claimed that visit time is a significant patients' satisfaction factor, even though it is considered indirectly.

Despite the importance of visit time in patient satisfaction, it may be sacrificed in walk-in clinics to gain more revenue. Furthermore, walk-in clinics benefit from the number of patients they serve and more admitted patients will result in increase in revenue. It would be better strategy to reduce visit time to reject less patients. However, in this situation we are witnessing a gap between what patients expect and what they receive. For this matter, Mostafa and Hamed [44] studied a walk-in clinic as a queuing system with assuming exponential distributions for time-between-arrival and service time, resulting in $M/M/1$ system. They concluded, government intervention and a regulation in the form of minimum visit time would be an effective way to increase patient satisfaction.

2.3 Research gaps

The research gaps that motivated this research are divided into three categories:

- Sometimes the arrival rate function is neither approximately linear nor approximately quadratic. Its behaviour is like a cubic function. In this case, we should use a polynomial function of degree 3 to approximate the arrival rate. To the best of the author's knowledge, a polynomial function of degree 3 has not been considered to approximate the arrival rate in a walk-in clinic.

- Long visit time leads to overcrowding in the uncapacitated clinics. So far, overcrowding has not been considered as an issue in walk in clinics with infinite capacity.

- Long visit time is also a problem in the capacitated clinics. Because if a clinic spends a long time on admitted patients, the departure rate would be slow, and due to not having enough space in the waiting room, the clinic has to reject new arriving patients. Any rejection means losing the revenue that could be obtained from that patient. This situation becomes more serious in rush hours. Since a walk-in clinic is established for business purposes, many capacitated clinics tend to allocate minimum visit time especially in peak hours. The suggested methods to solve this problem for protecting patients' rights are bound to government's intervention. However, incentive ways of prompting clinics to maintain the quality of service even in rush hours would be more efficient.

2.4 Contribution

This research offers a new analytical and methodological approach to fill the mentioned gaps:

- A new approximation in the form of polynomial function of degree 3 will be suggested to estimate arrival rates. The parameters of the suggested function are approximated by using ordinary least squares (OLS) method. Then waiting time in an ideal system with a cubic time-varying arrival rate will be estimated.

- Many researchers have considered the $M/M/k$ for walk-in clinics, while the model $M_t/M/k$

would be more appropriate in reality. Such model with linear, quadratic and cubic polynomial functions approximating the arrival rates will be considered for walk-in clinics.

- Overcrowding will be considered as the negative factor influencing patients' satisfaction in unlimited capacity clinics. With considering this factor, the optimum value maximizing patients' satisfaction is obtained by using numerical methods.

- Since capacitated walk-in clinics benefit from the number of patients they admit, it is in their best interest to allocate minimum visit times to acquire maximum revenue. Hence, quality of service may be sacrificed to obtain more revenue. In this research a new method is suggested which solves the problem by giving funds based on the performance of walk-in clinics. The method of rewarding increases patients' satisfaction and clinics' profit simultaneously.

3. Queuing theory with time-varying arrival rate

In this section, first a brief look will be taken into the formulation of queuing theory to provide the reader with enough background to properly model a queuing system. Then, the basic queuing model will be defined, and notations, queuing disciplines, birth-death processes, and Little's queuing formula will be discussed. Finally, queues with time-dependent parameters and estimation of waiting time will be reviewed.

3.1 Stochastic processes

To begin understanding queues, the readers must have some knowledge of probability theory. In particular, Markov chains, the exponential and Poisson probability distributions will be reviewed.

3.1.1 Markov chains

In probability theory and statistics, Markov chains are a common way to model random processes. They have been used in many different domains, ranging from text generation to financial modeling. A Markov chain is a mathematical system experiencing transitions from one state to another according to certain probabilistic rules. It is a stochastic process, but what differentiates a Markov chain from a general stochastic process is Its "memory-less" property. Regardless of how the process arrived at its present state, the possible future states are fixed. In other words, the probability of transitioning to any particular state is dependent only on the current state. Consider a sequence $X_0, X_1, ...$ of random variables satisfying the rule of conditional independence. It is a Markov chain if for any positive integer n and possible states $i_0, i_1, ..., i_n$ of the random variables, it satisfies:

$$P(X_n = i_n | X_0 = i_0, X_1 = i_1, ..., X_{n-1} = i_{n-1}) = P(X_n = i_n | X_{n-1} = i_{n-1}).$$

3.1.2 The exponential distribution

A continuous random variable X is said to have an exponential distribution with parameter λ if its PDF is given by

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

It is shown by:

$$X \sim Exp(\frac{1}{\lambda}).$$

If X is a random variable that represents inter-arrival times with the exponential distribution, then $P(X \leq x) = 1 - e^{-\lambda x}$. The most important property of the exponential distribution is its memory-less property. This can be formally stated for all non-negative values of t and h as follows:

$$P(X > t + h|A \geq t) = P(X > h).$$

Due to having the no-memory property, the distribution lends itself well to modeling customer inter-arrival times or service times. The no-memory property suggests that the time until the next arrival will never depend on how much time has already passed. This makes intuitive sense for a model where customer arrivals are being measuring, because the customers' actions are clearly independent of one another.

3.1.3 The Poisson process

The Poisson process is the canonical traffic process model. In addition, the Poisson distribution is used to determine the probability of a certain number of arrivals occurring in a given time period. For instance, a call center receives an average of 60 calls per hour. The calls are independent which means receiving one does not affect the probability of next call. The number of calls received during any minute has a Poisson probability distribution: the most likely numbers are 0,1 and 2, while there is a very small probability it could be 7.

The Poisson distribution with parameter λ is given by

$$f_X(x) = \begin{cases} \frac{(\lambda)^x e^{-\lambda}}{x!} & x = 0, 1, 2, ... \\ 0 & \text{otherwise} \end{cases}$$

where x is the number of arrivals. A Poisson random variable is represented by:

$$X \sim Po(\lambda).$$

The relationship between Poisson and exponential distribution

It is useful to note the exponential distribution's relation to the Poisson distribution. If X shows the number of events which are likely to occur in the interval of time $[0, T]$, and T represents the expected time for the next event, then:

$$X \sim Po(\lambda) \Rightarrow T \sim Exp(\frac{1}{\lambda}).$$

In addition, in a Poisson process, if events accrue on average at the rate λ per unit of time, then there will be on average λt occurrence per t units of time. The Poisson distribution describing this process is

$$P(x) = \frac{(\lambda t)^x e^{-\lambda t}}{x!}$$

where $P(x = 0) = e^{-\lambda t}$ is the probability of no occurrences in t units of time. Also, $P(x = 0) = e^{-\lambda t}$ shows the probability that the time, T, to the first occurrence is greater than t:

$$P(T > t) = P(x = 0|\mu = \lambda t) = e^{-\lambda t}$$

In contrast, the probability that an event does occur during t units of time is shown by

$$P(T \leq t) = 1 - P(x = 0|\mu = \lambda t) = 1 - e^{-\lambda t}$$

As shown, this is the cumulative exponential distribution. If it is differentiated with respect to t, produces the probability density function of the exponential distribution

$$f_T(t) = \lambda e^{-\lambda t}$$

3.2 Queuing modelling fundamentals

This section explains queuing modelling fundamentals in this section. The six fundamental elements are:

- The arrival process,
- The service process,
- The number of servers,
- The queuing discipline,
- The queue capacity, and
- The population.

First, each element will be explained and then Kendall's notation, which is the standard system used to describe and classify a queuing model, will be introduced.

3.2.1 Arrival process

Arrival represents the way customers enter the system. In usual queuing situations, the process of arrival in the system is stochastic because at a given period customers arrive randomly. The arrival of one customer is also independent of arrival of another one. In the fallowing, a call center with *made up data* is considered to clarify the process of calculating the expected arrival rate.

Let us to consider a call center and observe the number of calls received over a 24-hours period. During 24 hours of observational survey, a period, for example 10 minutes, should be defined for one slot of observation. Then, an observation line far behind the ordinary waiting line to count the number of new arriving calls within 10 minutes observation should be set. Table 3.1 presents data taken from the observation of a call center within a 24-hour interval.

Table 3.1: Time and number of arrival

Time	Number of arrival
00:00-00:10	1
00:10-00:20	3
...	...
23:40-23:50	2
23:50-00:00	0
Total	544

Using the number of arrivals, arrival distribution can be drawn. Arrival distribution shows how many times a certain number of customers arrive within 10 minutes are observed. Table 3.2 shows arrivals in a call center.

Table 3.2: Arrivals in a call center

Number of arrivals	Count	Relative frequency
0	260	48%
1	186	34%
2	75	14%
3	18	3%
4	5	1%
Total	544	100%

The arrival rate in that call center is computed as follows:

$$\lambda = 0 * 48\% + 1 * 34\% + 2 * 14\% + 3 * 3\% + 4 * 1\% = 0.75,$$

or 0.75 call is expected per 10 minutes. The most common type of arrival distribution in a queuing system follow Poisson process. As discussed, the Poisson distribution with parameter λ is given by

$$P(n) = \frac{\lambda^n e^{-\lambda}}{n!}, \quad n = 0, 1, 2, ...$$

Table 3.3 shows the probability based on Poisson formula:

Table 3.3: Poisson distribution of arrivals

Number of arrivals, n	Poisson probability $P(n)$ with $\lambda = 0.75$
0	47.24%
1	35.43%
2	13.29%
3	3.32%
4	0.62%
5 0r more	0.1%

As shown in Tables 3.2 and 3.3, probability of Poisson distribution and relative frequency of observations are almost similar. If t_i is defined as the time when the ith customer arrives, then $T_i = t_{i+1} - t_i$ would be the ith interarrival time. It is usually assumed that all T_i's are independent, continuous and random variables. As discussed in section 3.1.3, if the number of arrival in a given period of time occurs randomly and independently from other arrivals and follows a Poison distribution with mean λ, then the inter-arrival time distribution follows an exponential probability distribution with mean $\frac{1}{\lambda}$.

3.2.2 Service process

To obtain distribution of service times, first data for each server over a given period of observation time is collected. Then, the time at which a customer begins to be served and the time that s/he has been served and left the line is recorded.

Table 3.4: Frequency of service times

service time	Count	Relative frequency	Cumulative frequency
0.5	261	48%	48%
1	129	24%	72%
1.5	75	14%	86%
2	35	6.5%	92.5%
2.5	17	3%	95.5%
3	15	2.5%	98%
3.5	5	1%	99%
4	3	0.5%	99.5%
4.5	0	0%	99.5%
5	2	0.3%	99.8%
5.5	1	0.1%	99.9%
6	0	0.0%	99.9%
6.5	1	0.1%	100%
Total	544	100%	

Table 3.4 shows the observed data for frequency of service times in a call center. To obtain service time, arriving and departure times are calculated. In addition, consider the starting time to serve a customer is 08:05:35 and the completion time is 08:07:06. Therefore, the service time shown by $t_{service}$ is 01:31 (Minute: Second).

Data are collected and observed to obtain the average service time $E[t_{service}]$. Then the service rate which is equal to

$$\mu = \frac{1}{E[t_{service}]} \tag{3.1}$$

is obtained. Service rate represents the average number of customers being served per unit of time. For instance, if the average service time is

$$E[t_{service}] = 1.08,$$

then the average number of customers being served per minute will be

$$\mu = \frac{1}{1.08} = 0.926.$$

When the the results shown in Table 3.4 is compared to a theoretical distribution, surprisingly it is found that cumulative distribution is close to the cumulative exponential distribution. Hence, the probability that the service time is less than or equal to a time length x is given by

$$P(t_{service} < x) = 1 - e^{-\mu x}.$$

A vast majority of service distributions in a queuing system follow exponential process. Table 3.5 represents similarity of theoretical distribution to the observed distribution.

Table 3.5: Exponential distribution of service times

service time	Cumulative frequency	Theoretical cumulative distribution
0.5	48%	41.7%
1	72%	66.01%
1.5	86%	80.18%
2	92.5%	88.44%
2.5	95.5%	93.26%
3	98%	96.07%
3.5	99%	97.71%
4	99.5%	98.66%
4.5	99.5%	99.22%
5	99.8%	99.55%
5.5	99.9%	99.74%
6	99.9%	99.85%
6.5	100%	99.91%
Total	100%	99.91%

3.2.3 The number of servers

Idle time in queuing systems occurs whenever a server is not busy. A server being in such state is referred to as an available (free) server. It is common to assume that an available server is willing and able to start serving whenever there is a demand for service. Generally, a system has finite number of servers. If a new customer arrives while all servers are busy, the customer has to wait in queue for the next available server.

3.2.4 The queuing discipline

The method in which arrivals in a queue get processed is known as the queuing discipline. The discipline determines the rule to select the next customer. It is easy for one to think of all queues operating like a grocery checkout line. When an arrival occurs, it is added to the end of the queue and service is not performed on it until all arrivals, that came before it, are served in the order they arrived. Although this a very common method for queues to be handled, it is far from the only way. This particular example outlines a First-Come-First-Serve discipline. However, different types of laws are used in different situation. The most commonly used laws are:

- FIFO - First In First Out: who comes earlier leaves earlier, FCFS - First Come First Served
- LIFO - Last Come First Out: who comes later leaves earlier, LCFS - Last Come First Served
- RS - Random Service: the customer is selected randomly, SIRO - Service In Random Order
- Priority without Preemption or Head of Line (HOL), Priority with Preemption, Resume or Repeat
- PS - Processor Sharing.

3.2.5 The queue capacity

The capacity of a queue is the number of elements the queue can hold. In other words, a queuing system has a space in which a limited number of customers can be accepted. Therefore, when the space is full, no customer is accepted. However, there are some systems which don't have any threshold for admitting customers. In this case, the capacity is considered to be infinite.

3.2.6 The calling population

The population of potential customers is named the calling population which can be finite or infinite. The key difference between finite and infinite population model is how the arrival rate is defined:

- Finite population model: if arrival rate depends on the number of customers being served and waiting

- Infinite population model: if arrival rate is not affected by the number of customers being served and waiting

The figure below shows a queuing system:

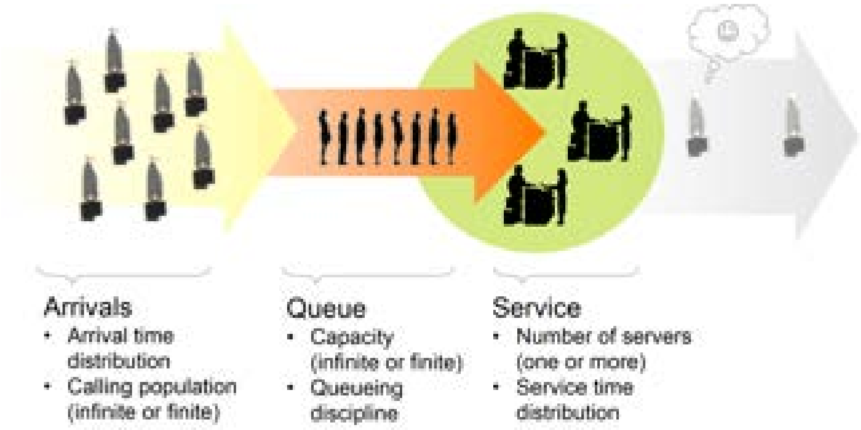
Arrivals
• Arrival time distribution
• Calling population (infinite or finite)
Queue
• Capacity (infinite or finite)
• Queueing discipline
Service
• Number of servers (one or more)
• Service time distribution

Figure 3.1: Queuing System

3.2.7 Kendall's notation

Since describing all characteristics of a queue inevitably becomes very wordy, a much simpler notation, known as Kendall-Lee notation, is used to describe a system. Kendall-Lee notation gives us six abbreviations for characteristics listed in order separated by slashes:

$$A/B/m/K/n/D$$

where the notations are defined as bellows:

A: distribution function of the inter-arrival times,

B: distribution function of the service times,

m: number of servers,

K: capacity of the system, the maximum number of customers in the system including the one being serviced,

n: population size, number of sources of customers,

D: service discipline.

Exponentially distributed random variables are notated by M, meaning Markovian or memoryless. Hence, $M/M/1$ denotes a system with Poisson arrivals, exponentially distributed service times and a single server. $M/G/m$ denotes an m-server system with Poisson arrivals and generally distributed service times. $M/M/r/K/n$ stands for a system where the customers arrive from a finite-source with n elements, inter-arrival and service times are exponentially distributed, the service is carried out according to the request's arrival by r severs, and the system capacity is K.

The aim of all investigations in queuing theory is to get the main performance measures of the system which are the probabilistic properties (distribution function, density function, mean, variance) of the following random variables: number of customers in the system, number of waiting customers, utilization of the server/s, response time of a customer, waiting time of a customer, idle time of the server, busy time of a server. Of course, the answers heavily depend on the assumptions concerning the distribution of inter-arrival times, service times, number of servers, capacity, and service discipline. It is quite rare, except for elementary or Markovian systems, that the distributions can be computed. Usually, their mean or transforms can be calculated.

3.2.8 Utilization factor

Apart from the mentioned parameters, there is another significant parameter in queuing systems named "loading factor of the service server" or "utilization factor,". It is denoted with ρ and defined as the portion of time the service station is busy and cannot serve other customers. Since the system cannot perform more work than its capacity allows for, the upper bound of the utilization factor is restricted by Petrovic et al. [45]

$$\rho = \{\frac{\lambda}{m\mu}, 1\}$$

where m is the number of servers. In an IS models, it is assumed that:

$$\rho \approx 0.$$

Also, in a stable system

$$\rho \leq 1. \tag{3.2}$$

$\rho > 1$ means that more customers arrive to the system than exit. Then the length of queue tends to increase to infinity. Such systems are called unstable systems. Usually performance of stable systems in which $\rho \leq 1$ is investigated.

3.3 Birth-death process

The birth–death processes (BDPs) are a flexible class of continuous-time Markov chains which are used to model the number of "particles" in a system, where each particle can "give birth" to another particle or "die". In addition, a BDP is a continuous-time Markov chain $X(t)$ counting the number of particles in a system at time t, taking values on the non-negative integers N. The model's name comes from the science of biology, since biologists study the development of populations of organisms by using the birth-death process.

For constructing a general BDP in a formal way, the rules according to which the number of particles evolves must be defined. This is done by specifying the behavior of the process for a very short time dt, when there are n particles in the system. If dt is very small, the probability of an event in the interval $(t, t + dt)$ that occurs with rate r is approximately rdt. Hence, the probability of a birth in the interval $(t, t + dt)$, given $X(t) = k$, is

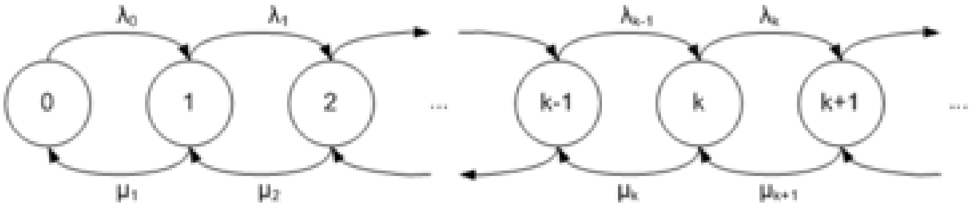

Figure 3.2: The birth death Markov processes

$$P(X(t + dt) = k + 1|X(t) = k) = \lambda_k dt + o(dt) \tag{3.3}$$

As shown, the probability of more than one birth event in a small time dt is significantly small. Also, the probability of a death in $(t, t + dt)$ is

$$P(X(t + dt) = k - 1|X(t) = k) = \mu_k dt + o(dt) \tag{3.4}$$

In addition, The probability of no births or deaths occurring during $(t, t + dt)$ is

$$P(X(t + dt) = k|X(t) = k) = 1 - (\lambda_k + \mu_k)dt + o(dt) \tag{3.5}$$

3.3.1 Transition probabilities

Consider $P_{ab}(t) = P(X(t) = b|X(0) = a)$ represents transition probability from state $X(0) = a$ to $X(t) = b$. Suppose that $X(0) = a$. At the current time t, we want to know the probability that in the next dt units of time, the process will reach state b. We look into the future by writing the probabilities of three types of events that can take the process to state b:

- Birth from $b - 1$
- Death from $b + 1$ or
- No change from b.

So, the result is:

$$P_{ab}(t + dt) = \lambda_{b-1} P_{a,b-1}(t)dt + \mu_{b+1} P_{a,b+1}(t)dt + (1 - \lambda_b - \mu_b) P_{ab}(t)dt + o(dt)$$

Subtracting $P_{ab}(t)$ from both sides, dividing by dt, and sending dt to zero, the Kolmogorov forward equation is obtained:

$$\frac{dP_{ab}(t)}{dt} = \lambda_{b-1} P_{a,b-1}(t) + \mu_{b+1} P_{a,b+1}(t) - (\lambda_b + \mu_b) P_{ab}(t) \tag{3.6}$$

where $P_{ab}(0) = 1$ if $a = b$ and zero otherwise.

3.3.2 Equilibrium probability

While calculating equilibrium probabilities for a Markov process, it is assumed that transition probabilities do not change. Let's set the left-hand side of the Kolmogorov forward equation (3.6) to zero and replace the finite-time transition probabilities $P_{ab}(t)$ with the equilibrium probabilities ($P_{ab} \to \pi_b$) to get:

$$0 = \lambda_{b-1} \pi_{b-1} + \mu_{b+1} \pi_{b+1} - (\lambda_b + \mu_b) \pi_b$$

resulting the equation

$$\mu_{b+1} \pi_{b+1} - \lambda_b \pi_b = \mu_b \pi_b - \lambda_{b-1} \pi_{b-1}. \tag{3.7}$$

Since this is true for every b, it is also true for $b = 0$. It is usually considered that $\mu_0 = \pi_{-1} = \lambda_{-1} = 0$, so both sides of (3.7) are zero for every b by induction. This gives the detailed balance condition for continuous-time Markov chains,

$$\mu_i \pi_i = \lambda_{i-1} \pi_{i-1} \qquad i = 0, 1, 2, ... \tag{3.8}$$

Therefore, every general BDP is a reversible Markov chain. Iterating the recurrence (3.8), it is found that

$$\pi_i = \frac{\lambda_0 \lambda_1 \lambda_2 ... \lambda_{i-1}}{\mu_1 \mu_2 \mu_3 ... \mu_i} \pi_0. \tag{3.9}$$

3.4 The non-homogeneous Poisson process

The Poisson process is a counting process for the number of events that occur at a certain time, with a parameter λ. The Poisson process is a special activity of the process of counting where intervals of events are mutually independent, have free increases and all are exponentially distributed. If the exponential distribution has the same parameter value, then it is called a homogeneous Poisson process. However, if it is not the same, it is called a non-homogeneous Poisson process. In addition, non-homogeneous Poisson processes are Poisson processes with parameters that depend on time and are not constant from time to time, they are also mutually independent. The emergence of new services and new technologies has led to new traffic models. Al- though there are some situations in which the Poisson process traffic model is still appropriate, in many new situations the classical Poisson process traffic model is not. In these situation, non- homogeneous Poisson processes can be applied. For example, these days for measuring daily ozone gas, model of noise exposure, and new approaches to improving software reliability models the non-homogeneous Poisson processes is used.

In reality the arrival process is also non-homogeneous and arrival rate typically varies significantly in time. In this section, a non-homogeneous Poisson process will be reviewed and ways to estimate its parameters with linear, quadratic and cubic function over a finite interval will be investigated.

Non-homogeneous Poisson processes (NHPPs) are widely used to model time-dependent arrivals in a multitude of stochastic models. Their widespread use is because of the fact that they may be defined in terms of very natural assumptions about the mechanism through which events happen. In particular, when customers arrive somewhat randomly to the system, the number of customers arriving to the system is modeled as a Poisson process with a non-stationary rate. Moreover, if $N(t)$

represents the number of arriving customers by time t, it can be modeled as a non-homogeneous Poisson process. Such process has all the properties of a Poisson process, except for the fact that its rate is a function of time, i.e., $\lambda = \lambda(t)$. The issue is estimating the rate function. A well-known heuristic for estimating the rate function of a non-homogeneous Poisson process assumes that the rate function is piece-wise constant on a set of data-independent intervals. It will be reviewed in more detail below. First, let's define a non-homogeneous Poisson process mathematically.

3.4.1 Definition

A counting process $N(t), t \geq 0$ is called a non-homogeneous Poisson process if:

- for $t, s \geq 0$, and $0 \leq u \leq t$, $N(t+s) - N(t)$ is independent of $N(u)$;
- for $t, s \geq 0$, $Pr(N(t+s) - N(t) \geq 2) = o(s)$
- for $t, s \geq 0$, $Pr(N(t+s) - N(t) = 1) = \lambda(t)s + o(s)$.

The function $\lambda(t)$ appearing in the definition is called the rate function which characterizes the Poisson process. Also, the notation $o(s)$ is used in the usual sense to denote a function $f(s)$ that satisfies $\lim_{s \to 0} \frac{f(s)}{s} = 0$.

3.4.2 Characteristic properties

Some characteristic properties of a non-homogeneous Poisson process are:

- The number of points in any interval has a Poisson distribution.
- The number of points in any finite set of non overlapping intervals are mutually independent random variables.
- The intervals between the points are not independent.
- The intervals between the points are not identically distributed.

The most general NHPP is defined in terms of a monotone non decreasing right-continuous function $\Lambda(t)$ bounded in any finite interval:

$$\Lambda(t) = \int_0^t \lambda(s)ds. \tag{3.10}$$

.

Then the number of points in any finite interval, for example $(0, t_0]$ has a Poisson distribution with parameter $\mu = \Lambda(t_0) - \Lambda(0)$. The right derivative of $\Lambda(t)$ is $\lambda(t)$ which is called rate function. $\Lambda(t)$ is called the integrated rate function and has the interpretation that

$$E[N(t)] = \Lambda(t) - \Lambda(0) = \int_0^t \lambda(s)ds, \tag{3.11}$$

where $N(t)$ indicates the total number of points in $(0, t]$.

3.5 Estimating the parameters of a non-homogeneous Poisson process

A non-homogeneous Poisson process model is parameterized by its arrival rate function $\lambda(t)$. In many cases it is reasonable to regard the arrival-rate function as linear, quadratic or cubic over appropriate sub-intervals. In this research, arrival rates will be estimated first by a linear function and then by quadratic and cubic function.

Consider a non-homogeneous Poisson process over the interval $[0, T]$ with linear arrival rate function:

$$\lambda(t) = a + bt, \qquad 0 \leq t \leq T. \tag{3.12}$$

Now two parameters a and b should be estimated. Assume the overall time interval $(0, T]$ is divided into N measurement sub-intervals

$$(\frac{(k-1)T}{N}, \frac{kT}{N}], \qquad 1 \leq k \leq N$$

and then observe the number of points in each. The estimation is based on a single realization of an arrival process or multiple independent samples. Massey at el. [13] investigated ways to estimate the parameters of a non-Homogeneous Poisson Process with linear rate over a finite interval. They considered:

- Ordinary Least Squares (OLS),
- Iterative Weighted Least Squares (IWLS) and
- Maximum Likelihood (ML) methods.

When the rate function is not near 0 at either end, approximately the same results will be obtained and none of the procedures differ considerably. In this research, we opt for the ordinary least squares estimators.

3.5.1 The ordinary least squares estimators

Consider a non-homogeneous Poisson process with linear rate

$$\lambda(t) = a + bt, \qquad 0 \leq t \leq T,$$

then count the number of points in the N sub-intervals

$$(\frac{(k-1)T}{N}, \frac{kT}{N}], \qquad 1 \leq k \leq N.$$

This sampling procedure from a single realization of the non-homogeneous Poisson Process over $[0, T]$ produce mutually independent Poisson random variables Y_k with mean

$$\lambda_k = \frac{T}{N}(a + bx_k) \tag{3.13}$$

where

$$x_k = (k - \frac{1}{2})\frac{T}{N}. \tag{3.14}$$

Because if the total number of points in each sub-interval is shown with $N(((k-1)\frac{T}{N}, k\frac{T}{N}])$, the mean is:

$$E[N(((k-1)\frac{T}{N}, k\frac{T}{N}])] = \Lambda(k\frac{T}{N}) - \Lambda((k-1)\frac{T}{N}) = \int_{(k-1)\frac{T}{N}}^{k\frac{T}{N}} \lambda(s)ds = \int_{(k-1)\frac{T}{N}}^{k\frac{T}{N}} (a + bs)ds$$

$$= \frac{T}{N}(a + b((k - \frac{1}{2})\frac{T}{N})) = \frac{T}{N}(a + bx_k).$$

If we form the linear model

$$Y = \beta_0 + \beta_1 x + \epsilon$$

and assume

$$y_k = \beta_0 + \beta_1 x_k + \epsilon_k, \qquad 1 \leq k \leq N,$$

then, parameters β_0 and β_1 and consequently parameters a and b can be approximated. To do that, sum of the squared errors should be minimized:

$$min \sum_{k=1}^{N} {\epsilon_k}^2 = \sum_{k=1}^{N} (y_k - [\beta_0 + \beta_1 x_k])^2. \tag{3.15}$$

Applying calculus with (3.15) in the usual way, $\hat{\beta_0}$ and $\hat{\beta_1}$ are obtained as bellow:

$$\hat{\beta_1} = \frac{\sum_{k=1}^{N}(x_k - \overline{x})(y_k - \overline{y})}{\sum_{k=1}^{N}(x_k - \overline{x})^2} \tag{3.16}$$

$$\hat{\beta_0} = \overline{y} - \hat{\beta_1}\overline{x} \tag{3.17}$$

where

$$\overline{y} = \frac{\sum_{k=1}^{N} y_k}{N}$$

$$\overline{x} = \frac{\sum_{k=1}^{N} x_k}{N}.$$

From (3.13), it can be seen that

$$\beta_0 = \frac{aT}{N}$$

$$\beta_1 = \frac{bT}{N}.$$

Using $\hat{\beta_0}$ and $\hat{\beta_1}$ as the estimation of β_0 and β_1 respectively, $\hat{a}$ and $\hat{b}$ are obtained:

$$\hat{a} = \frac{N}{T}\hat{\beta_0} \tag{3.18}$$

$$\hat{b} = \frac{N}{T}\hat{\beta_1}. \tag{3.19}$$

The resulting estimators are unbiased:

$$E[\hat{a}] = a$$

$$E[\hat{b}] = b.$$

Time varying arrival rates would be approximated by polynomial function of degree p such that $p \geq 2$. To find the coefficients of polynomial function, a Multiple Linear Regression model can be applied. Moreover, consider a single dependent variable y and several independent variables $x_1, x_2, ..., x_p$. In Multiple Linear Regression, the following model is assumed:

$$y = \beta_0 + \beta_1 x_1, ..., \beta_p x_p + \epsilon \tag{3.20}$$

where $\beta_0, \beta_1, ..., \beta_p$ are unknown parameters of the function f and ϵ is a random disturbance (usually assumed to have a normal distribution with mean 0 and standard deviation σ).

Suppose N observations for dependent and independent variables have been given as shown in matrix Y and matrix X:

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$

$$X = \begin{bmatrix} x_{11} & x_{12} & ... & x_{1p} \\ x_{21} & x_{22} & ... & x_{2p} \\ \vdots & & & \\ x_{N1} & x_{N2} & ... & x_{Np} \end{bmatrix}.$$

The coefficients $\{\beta_j\}_{j=0}^{p}$ which fit the equations best, can be found by solving the minimization problem

$$\hat{\beta} = \arg\min_{\beta} S(\beta)$$

where

$$S(\beta) = \|y - X\beta\|^2 = \Sigma_{k=1}^{N} |y_k - (\beta_0 + \beta_1 x_{k1} + \beta_2 x_{k2} + ... + \beta_p x_{kp})|^2.$$

In general, assume a non-homogeneous Poisson process with a polynomial function rate of degree p:

$$\lambda(t) = a_0 + a_1 t + a_2 t^2 + ... + a_p t^p, \qquad 0 \le t \le T.$$

To estimate the parameters $\{a_j\}_{j=0}^{p}$, we consider $y = \lambda(t)$ and $x_k = t^k, k = 1, ..., N$ and then use the regression methods as mentioned above.

3.6 Perturbation theory

Generally, finding the close-form solution of many problems is impossible or at least so difficult such that it is not practical to obtain. However, sometimes it would be possible to obtain a so-called asymptotic series approximation of the solution giving a good estimation to the solution. In this section, perturbation methods, which are used to obtain approximate analytic solutions to polynomials that can't be solved exactly, will be reviewed.

3.6.1 Preliminary material

In this method, the expansion of the algebraic expression $(x+y)^n$ will be required. The binomial theorem can be applied to state the expression as a sum of the terms involving individual exponents of variables x and y. According to the theorem, it is possible to expand any nonnegative integer power of $(x + y)$ into a sum of the form

$$(x + y)^n = \begin{pmatrix} 0 \\ n \end{pmatrix} x^0 y^n + \begin{pmatrix} 1 \\ n \end{pmatrix} x^1 y^{n-1} + \begin{pmatrix} 2 \\ n \end{pmatrix} x^2 y^{n-2} + ...$$

When n is a positive integer, this formula terminates:

$$(x + y)^n = \begin{pmatrix} 0 \\ n \end{pmatrix} x^0 y^n + \begin{pmatrix} 1 \\ n \end{pmatrix} x^1 y^{n-1} + \begin{pmatrix} 2 \\ n \end{pmatrix} x^2 y^{n-2} + ... + \begin{pmatrix} n \\ 0 \end{pmatrix} x^n y^0.$$

Another useful theorem in this method is the fundamental theorem of perturbation theory. Before expressing it, an asymptotic expansion should be defined.

Definition of an asymptotic expansion: The series

$$\Sigma_{n=0}^{N} c_n x_n(\epsilon)$$

is an asymptotic expansion of $x(\epsilon)$ at $\epsilon = 0$ if the following hold:

- $x_n(\epsilon) = o(x_{n-1}(\epsilon))$ as $\epsilon \rightarrow 0$ for $n = 1, 2, ..., N + 1$
- $x(\epsilon) - \Sigma_{n=0}^{N} c_n x_n(\epsilon) = O(x_{N+1}(\epsilon))$ as $\epsilon \rightarrow 0$

Theorem 3.6.1. *If an asymptotic expansion satisfies*

$$A_0 + A_1\epsilon + ... + A_N\epsilon^N + O(\epsilon^{N+1}) = 0$$

for all sufficiently small ϵ and the coefficients A_j are independent of ϵ, we have

$$A_0 = A_1 = ... = A_N = 0$$

3.6.2 Description of perturbation theory

In perturbation method, the problem is divided into ”solvable” and ”perturbative” parts. First, the solvable part is considered in order to find an estimation of the solution, then it works for continuously improving the previously obtained approximation. Assume the problem is finding the roots of a polynomial of degree n, so the perturbed equation is

$$f(x) = c_0 + c_1 + ... + c_n x^n = 0.$$

Consider an approximation for the full solution $x(\epsilon)$ which is an asymptotic series in the small parameter ϵ, like the following:

$$x(\epsilon) = a_0 + a_1\epsilon + a_2\epsilon^2 + ...$$

where a_0 is known solution and $\{a_i\}_{i=1,2,...}$ can be found iteratively by a mechanistic procedure. Furthermore, this formal series is substituted into the perturbed equation and then set the terms corresponding to powers of ϵ equal to zero and try to find the $\{a_i\}_{i=1,2,...}$. An approximate perturbative solution is obtained by truncating the series, often by keeping only the first two terms,

$$x(\epsilon) = a_0 + a_1\epsilon + O(\epsilon),$$

where $O(\epsilon)$ indicates the order of the error in the approximate solution.

3.6.3 Algebraic equations

Example: Consider the cubic equation

$$x^3 + 0.01x + 27 = 0.$$

We probably do not know how to solve this equation to find the exact solution. In this case, 0.01 is small compared to other coefficients. This suggests that the equation studied is:

$$x^3 + \epsilon x + 27 = 0. \tag{3.21}$$

It is expected that the root would be close to $x = -3$ which is the root of $x^3 + 27 = 0$. To find a better approximation, assume there is an asymptotic series in the form

$$x(\epsilon) = a_0 + a_1\epsilon + a_2\epsilon^2 + ...$$

Substitute this formal series into the cubic equation (3.21)

$$(a_0 + a_1\epsilon + a_2\epsilon^2 + ...)^3 + \epsilon(a_0 + a_1\epsilon + a_2\epsilon^2 + ...) + 27 = 0.$$

Collecting powers of ϵ leads to

$$(a_0^3 + 3a_0^2a_1\epsilon + 3a_0a_1^2\epsilon^2 + a_1^3\epsilon^3 + ...) + \epsilon(a_0 + a_1\epsilon + a_2\epsilon^2 + ...) + 27 = 0$$

or

$$(a_0^3 + 27) + \epsilon(3a_0^2a_1 + a_0) + ... = 0$$

By using fundamental theorem of perturbation theory (Theorem 3.1):

$$a_0^3 + 27 = 0 \Rightarrow a_0 = -3$$

$$3a_0^2a_1 + a_0 = 0 \Rightarrow a_1 = \frac{-a_0}{3a_0^2} = \frac{3}{27} = \frac{1}{9}$$

$$\vdots$$

Therefore, it is obtained

$$x(\epsilon) = a_0 + a_1\epsilon + a_2\epsilon^2 + ... = -3 + \frac{1}{9}\epsilon + ...$$

Since we want to keep just the two first terms, we get

$$x(\epsilon) = -3 + \frac{1}{9}\epsilon + O(\epsilon^2).$$

As a result, the approximation for the solution is

$$-3 + \frac{1}{9}(0.01) = -2.99888.$$

3.7 Statistical analysis with Little's Law

In queuing theory, an important discipline within the mathematical theory of probability is a theorem by John Little [17] that states that the long-term average number of customers in a stationary system is equal to the long-term average effective arrival rate multiplied by the average time that a customer spends in the system. In the following, we explain Little's Law algebraically.

3.7.1 Little's Law

Consider a queuing system in which customers arrive from the outside, spend time in the system and then depart. Let A_k be the arrival time and D_k be the departure time for customer k. we define

$$W_k \equiv D_k - A_k$$

Let $T_k(t)$ be the time of the kth arrival before time t (less than or equal to t) and $N(t)$ be the number of arrivals by time t:

$$N(t) = max\{k : T_k(t) \leq t\}.$$

Consider $L(t)$ as the total number of customers in the system at time t. Define

$$\lambda \equiv \lim_{t\to\infty} \frac{N(t)}{t}$$

$$W \equiv \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} W_k$$

$$L \equiv \lim_{t\to\infty} \frac{1}{t} \int_0^t L(s)ds$$

Theorem 3.7.1. *If both λ and W exist and are finite, then L exists and $L = \lambda W$*

Proof: See [17]

As shown, the theorem concerns either long-run averages (limits) or the expected values of stationary stochastic processes in stochastic models. Hence, the result does not necessarily directly apply over finite-time intervals, due to problems such as how to log customers already present at the start of the logging interval and those who have not yet departed when logging stops [28].

3.7.2 Measurements over a finite time interval

Kim and Whitt [28] investigated how to take a statistical approach with data over a finite-time interval. They assumed that the system was in operation in the past, prior to time 0, and that it will remain in operation after time t.

Let $R(0)$ count the customers that arrived before time 0 that remain in the system at time 0, so

$$L(0) = R(0) + N(0)$$

where $N(0)$ is the number of new arrivals at time 0, if any. We will carefully distinguish between $L(0)$ and $R(0)$. The averages of $W(t)$, $L(t)$ and $\lambda(t)$ over the time interval $[0, t]$ can be obtained:

$$\overline{\lambda}(t) \equiv \frac{N(t)}{t} \tag{3.22}$$

$$\overline{L}(t) \equiv \frac{1}{t} \int_0^t L(s)ds \tag{3.23}$$

$$\overline{W}(t) \equiv \frac{1}{N(t)} \sum_{k=R(0)+1}^{R(0)+N(t)} W_k. \tag{3.24}$$

It can be seen that $\overline{\lambda}(t)$ and $\overline{L}(t)$ are time average and have been observed, while $\overline{W}(t)$ is a customer average and cannot be directly observed. Hence, indirect estimator should be used. Kim and Whitt [28] created a new alternative estimator exploiting from $L = \lambda W$ in the form of

$$\overline{W}_{L,\lambda}(t) \equiv \frac{\overline{L}(t)}{\overline{\lambda}(t)}, \tag{3.25}$$

where $\overline{W}_{L,\lambda}(t)$ would be considered a substitute for $\overline{W}(t)$. Now, the question is ”how are the averages $\overline{W}_{L,\lambda}(t)$ and $\overline{W}(t)$ related?” The answer depends on the start and end points:

- If the system starts and ends empty ($L(t) = R(0) = 0$):

$$\overline{W}(t) = \overline{W}_{L,\lambda}(t)$$

- If system does not start empty or end empty ($R(0) \neq 0$ or $L(t) \neq 0$):

$$\overline{W}(t) = \overline{W}_{L,\lambda}(t) - \frac{T_W^{(r)}(0) - T_W^{(r)}(t)}{N(t)}$$

  where $T_W^{(r)}(t)$ is the process recording the total residual waiting time of all customers in the system at time t, which typically is not known if the waiting times are not directly observed. $T_W^{(r)}(t)$ is defined as

$$T_W^{(r)}(t) \equiv \sum_{k=1}^{L(t)} W_k^{r,t}$$

  where $W_k^{r,t}$ is the remaining waiting time at time t for customer k in the system at time t.

The issue is finding estimation of residual waiting time $W_k^{r,t}$. Kim and Whitt [28] considered estimation in two cases: when the system is stationary and when it is not:

- When the system is assumed to be stationary:

$$E[T_W^{(r)}(0)] = E[T_W^{(r)}(t)].$$

  So, it is reasonable to use the indirect estimator $\overline{W}_{L,\lambda}(t)$.

- When the system is assumed to be non-stationary as commonly happens when the arrival rate is time-varying, Kim and Whitt [28] used $\overline{W}_{L,\lambda}(t)$ to estimate the residual waiting time. In addition, they assumed the waiting time distribution remains fixed throughout the measurement interval and the distribution of the waiting times is nearly exponential. The exponential distribution assumption was used to justify approximating the residual waiting time distribution for each customer in the system at time t by the ordinary waiting time distribution, which in turn is estimated by $\overline{W}_{L,\lambda}(t)$. They proposed the following refined estimator:

$$W_{L,\lambda,r}(t) \equiv \overline{W}_{L,\lambda}(t) - \frac{(R(0) - L(t))\overline{W}_{L,\lambda(t)}}{N(t)} = \overline{W}_{L,\lambda}(t)(1 - \frac{R(0) - L(t)}{N(t)}). \tag{3.26}$$

Moreover, there is some underlying stochastic model for which the mean $E[\overline{W}(t)]$ is well defined and $\overline{W}(t)$ can be regarded as an estimate of $E[\overline{W}(t)]$. Kim and Whitt [28] used $\overline{W}_{L,\lambda}(t)$ as an estimator of $E[\overline{W}(t)]$. They found the bias in $\overline{W}_{L,\lambda}(t)$ to be an estimator is $E[\Delta_W(t)]$, where

$$\Delta_W(t) = \overline{W}_{L,\lambda}(t) - \overline{W}(t).$$

They proved in [28]

$$E(\Delta_W(t)) = E(E(\Delta_W(t)) : \theta(t))$$

where

$$E[\Delta_W(t) : \theta(t)] \approx \frac{(R(0) - L(t))\overline{W}_{L,\lambda(t)}}{N(t)}$$

$$\theta(t) = (t, \overline{L}(t), \overline{\lambda}(t), R(0), L(t)).$$

Then, they obtained the new candidate refined estimator of $E[\overline{W}(t)]$, exploiting the observed vector $(\overline{L}(t), \overline{\lambda}(t), R(0), L(t))$

$$W_{L,\lambda,r}(t) \equiv \overline{W}_{L,\lambda}(t) - E[\Delta_W(t) : \theta(t)] \approx \overline{W}_{L,\lambda}(t)(1 - \frac{R(0) - L(t)}{N(t)})$$

3.8 Estimating waiting times with time-varying Little's Law (TVLL)

As mentioned, when waiting times cannot be observed directly, Little's Law is used to estimate the average waiting time by the average number in system divided by the average arrival rate. However, that simple indirect estimator tends to be biased, especially when the arrival rates are time-varying. The bias in that indirect estimator can be estimated by applying the time-varying Little's law (TVLL). Kim and Whitt [21] assumed that there are appropriate time-varying servers, so the waiting time distribution would not be time-varying even though the arrival rate is time-varying. They fitted a linear and quadratic function to the time-varying arrival data. When the arrival rate function is approximated with a polynomial function of degree n, the mean waiting time satisfies an equation of degree $n + 1$. The new estimators based on the TVLL are positive real root of that equation.

3.8.1 Time Varying Little's Law (TVLL)

The TVLL is a time-varying generalization of LL. The arrival rate over the interval $[0, t]$ is specified by requiring that

$$E[N(I)] = E[N([t_1, t_2])] \equiv \Lambda(t_1, t_2) = \int_{t_1}^{t_2} \lambda(s)ds, \qquad -\infty < t_1 < t_2 < \infty.$$

We assume that the conditional cumulative distribution function (cdf)

$$G_t(x) \equiv P(W(t) \leq x | \mathcal{N}_t), \qquad x \geq 0$$

of the waiting time (time in system) for a new arrival at time t, given that an arrival occurs at time t (the event $\mathcal{N}_t$) is well defined for all t. (The precise meaning of the cdf G_t is somewhat complicated, see [19] and [20]. Its precise meaning is not very important because based on time-varying staffing property, the cdf G_t will be assumed independent of t). Consequently $G_s^c(x)$ is:

$$G_s^c(x) \equiv P(W(t) > x|\mathcal{N}_s), \qquad x \geq 0.$$

$L(t)$ as the total number of customers in the system at time t can be expressed as an infinite sum of random variables or, equivalently, as an elementary stochastic integral via:

$$\begin{aligned} L(t) = \sum_{k=1}^{\infty} 1_{\{W_k(t) \geq t - T_k(t)\}} &= \sum_{k=1}^{\infty} 1_{\{W(T_k(t)) \geq t - W(T_k(t))\}} \\ &= \int_{-\infty}^{t} 1_{\{W(s) \geq t-s\}} dN(s). \end{aligned} \tag{3.27}$$

Taking expectations in (3.27), the TVLL is obtained:

Theorem 3.8.1. *(TVLL): Under the conditions above*

$$E\left[L(t)\right] = \int_{-\infty}^{t} G_s^c(t-s)\lambda(s)ds$$

Proof: See [19] and [20]

As shown, it is not immediately apparent how to apply the TVLL in Theorem (3.8.1) to estimate waiting times. Kim and Whit [21] considered two additional assumptions.

3.8.2 Assumptions

Two strong assumptions considered by Kim and Whitt [21] are:

- Waiting time distribution remains fixed throughout the measurement interval. In addition, they supposed the distribution of $W(t)$ is distributed as W independent of t
- The fixed waiting time W has a cdf that is known except for its mean. In addition, it was supposed that there is a specified cdf G with mean 1 such that $P(W \leq xE(W)) = G(x), x \geq 0$

The first assumption is achieved by using appropriate time-varying staffing. With appropriate service providers, the waiting times tend to not significantly exceed the service times. Hence, that it is reasonable to regard the waiting times as stationary over sub-intervals.

The time-varying staffing

In many applications, time-varying staffing assumptions is reasonable. In well-managed centers, waiting times usually remain approximately stationary, even though the arrival rate may be time-varying. The time-varying staffing is selected to stabilize the performance at typical performance levels. The following formula is the method of Feldman et al. [46] and Jennings et al. [47] for required servers at time t:

$$S_t = [m(t) + \beta \sqrt{m(t)}] \tag{3.28}$$

where

S_t is the number of servers at time t,

$m(t) \equiv E[L(t)]$ is the offered load,

$[x]$ is the least integer greater than or equal to x, and

β is called the quality-of-service (QoS) which is usually considered: $0, 1, 2$.

$\beta = 0$ and $\beta = 1$ produce typical performance, whereas $\beta = 2$ corresponding to high QoS, produces performance close to the IS model.

3.8.3 The TVLL with fixed waiting time distribution

Eick et al. [48] studied the physics of the $M_t \backslash G \backslash \infty$ and showed that $Q(t)$, which represents the number of busy servers in the system at time t, has Poisson distribution with mean

$$E[Q(t)] = E[\int_{t-s}^{t} \lambda(u)du] = E[\lambda(t - S_e)]E[S] \tag{3.29}$$

where S shows service time and S_e is a random variable with the associated stationary-exceed or equilibrium-residual-lifetime cdf

$$G_e(t) \equiv P(S_e \leq t) \equiv \frac{1}{E[S]} \int G^c(u)du$$

where G is the cdf of S and

$$G^c(t) = 1 - G(t).$$

Moments of S_e are related to moments of S by

$$E[S_e^k] = \frac{E[S^{k+1}]}{(k+1)E[S]}, \qquad k \geq 1.$$

The TVLL in Theorem (3.8.1) has important connections to IS queuing models. In addition, in IS models the number of customers in the system is equal to the number of busy servers and the waiting times coincide with the service times. TVLL can be regarded as part of the theory for IS models, because the abstract system can be regarded as a general IS model, if we simply call the waiting time as the service time in the IS model. Under the first assumption, the TVLL in Theorem (3.8.1) reduces to the corresponding $M_t \backslash G \backslash \infty$ IS formula in Theorem 1 of [48]

$$E[L(t)] = E[\lambda(t - W_e)]E[W] \tag{3.30}$$

where the W and W_e are random variables with the fixed waiting-time cdf and the associated stationary-excess cdf, that is

$$P(W_e \leq t) \equiv \frac{1}{E[W]} \int_0^x P(W_e > u)du \tag{3.31}$$

$$E[W_e^k] = \frac{E[W^{k+1}]}{(k+1)E[W]}, \qquad k \geq 1. \tag{3.32}$$

3.8.4 An approximating linear arrival rate

Despite the fact that arrival rate is a time-varying function, it is approximately linear over subintervals, for example for an hour or two. So, a linear function can be applied to estimate arrival rate function:

$$\lambda(s) \approx \lambda_l(s) = a + bs, \qquad 0 \leq s \leq t \tag{3.33}$$

where a and b are constants such that $\lambda_l(s) \geq 0$, with $[0, t]$ denoting the designated time interval. This approximation can be obtained through a Taylor series approximation [48]. Furthermore, an ordinary least square fit was used in [13]. Using equations (3.30), (3.31), and (3.32), we get the associated approximation for $E\left[L(t)\right]$:

$$\begin{aligned} E\left[L(t)\right] &\approx \lambda_l(t - E\left[W_e\right])E\left[W\right] = \lambda_l(t - \gamma_W^2 E\left[W\right])E\left[W\right] \\ &= (a + bt)E\left[W\right] - b\gamma_W^2 E\left[W\right]^2 \end{aligned} \tag{3.34}$$

where

$$\gamma_W^2 = (c_W^2 + 1)/2$$

$$c_W^2 = Var(W)/E\left[W\right]^2.$$

By integrating over $[0, t]$ and dividing by t in (3.34), the following result is obtained:

$$E\left[\overline{L}(t)\right] \equiv t^{-1} \int_0^t E\left[L(s)\right] ds \approx (a + b(\frac{t}{2}))E\left[W\right] - b\gamma_W^2 E\left[W\right]^2 \tag{3.35}$$

Consider estimations $\hat{a}$ of a, $\hat{b}$ of b and $\overline{L}(t)$ of $E\left[\overline{L}(t)\right]$ and define

$$x \equiv E\left[W\right],$$

$$\hat{\lambda}_l(t) \equiv (\hat{a} + \hat{b}(\frac{t}{2}))$$
$$\hat{\lambda}_l' \equiv \hat{b},$$

then plug in x, $\hat{\lambda}_l(t)$, and $\hat{\lambda}_l'$ into (3.35). A quadratic equation is obtained as bellow:

$$\gamma_W^2 \hat{\lambda}_l' x^2 - \hat{\lambda}_l(t)x + \overline{L}(t) = 0. \tag{3.36}$$

By solving the equation (3.36), we get a new refined estimator based on a linear approximation of the arrival rate function

$$\overline{W}_{L,\lambda,l}(t) \equiv x \equiv \frac{B \pm \sqrt{B^2 - 4C}}{2} \tag{3.37}$$

for

$$B \equiv \frac{\hat{\lambda}_l(t)}{\gamma_W^2 \hat{\lambda}_l'}$$
$$C \equiv \frac{\overline{L}(t)}{\gamma_W^2 \hat{\lambda}_l'}.$$

If $\left|\gamma_W^2 \hat{\lambda}_l'\right|$ is too small, there will be numerical instability, because in calculating B and C, we divide by $\gamma_W^2 \hat{\lambda}_l'$. In this case, an alternative estimator can be calculated by using perturbation theory.

Perturbation analysis with a linear arrival rate function

Before introducing another new estimator, we express an important proposition.

Proposition 3.8.1. *Consider the quadratic equation*

$$a_2x^2 - a_1x + a_0$$

with $a_1 > 0$ *and* $a_0 > 0$ *and let*

$$\epsilon(a_2) \equiv \frac{a_2a_0}{a_1^2}.$$

If $4\epsilon(a_2) < 1$*, then the equation has two positive real roots and the minimum positive root can be expressed as*

$$x = \frac{a_0}{a_1}(1 + \epsilon(a_2) + o(a_2)) \qquad as \qquad a_2 \to 0.$$

Proof: Apply the Taylor series expansion

$$\sqrt{x+\epsilon} = \sqrt{x} + \frac{\epsilon}{2\sqrt{x}} - \frac{\epsilon^2}{8x^{\frac{3}{2}}} + o(\epsilon^2) \qquad as \qquad \epsilon \to 0.$$

Based on Proposition 3.8.1, and assuming $\hat{\lambda}_l(t) = \hat{\lambda}(t)$, the minimum positive root of the quadratic equation in (3.36) can be approximated by the perturbation method:

$$\overline{W}_{L,\lambda,l,p}(t) \equiv \overline{W}_{L,\lambda}(t)(1 + \overline{W}_{L,\lambda}(t)(\frac{\gamma_W^2\hat{\lambda}_l'}{\hat{\lambda}(t)})) \tag{3.38}$$

In other words,

$$\overline{W}_{L,\lambda,l,p}(t) \equiv \omega(1 + \omega\delta) \tag{3.39}$$

where

$$\omega = \overline{W}_{L,\lambda}(t)$$

$$\delta = \frac{\gamma_W^2 \hat{\lambda}_l'}{\hat{\lambda}(t)}$$

The estimator $\overline{W}_{L,\lambda,l,p}(t)$ is preferred to the estimator $\overline{W}_{L,\lambda,l}(t)$ when $\gamma_W^2 \hat{\lambda}_l'$ is small.

3.8.5 Estimating $R(0) - L(t)$ in $\overline{W}_{L,\lambda,r}(t)$

In the equation (3.26), the term $R(0)-L(t)$ is seen. We might be unable to observe $R(0)$ and $L(t)$, because we only have available $\overline{L}(t)$ and arrival process data, and do not have a full observation of $L(s), 0 \leq s \leq t$. In this case, it is assumed that $L(0) = R(0)$ and an estimate of $E[L(0)] - E[L(t)]$ is used instead. In addition, We first fit the arrival rate function to a linear function, then use the equation (3.34) to estimate $E[L(0)] - E[L(t)]$. It is obtained:

$$E[L(0)] - E[L(t)] \approx -btE[W] \tag{3.40}$$

If $E[W]$ is approximated with $\overline{W}_{L,\lambda}(t)$ and it is considered $\hat{\lambda}' = \hat{b}$, then:

$$E[L(0)] - E[L(t)] \approx -\hat{b}t\overline{W}_{L,\lambda}(t) \approx -\hat{\lambda}' t\overline{W}_{L,\lambda}(t) \tag{3.41}$$

Therefore,

$$\begin{aligned} W_{L,\lambda,r}(t) \equiv \overline{W}_{L,\lambda}(t)(1 - \frac{R(0) - L(t)}{N(t)}) &\approx \overline{W}_{L,\lambda}(t)(1 - \frac{E[L(0)] - E[L(t)]}{N(t)}) \\ &\approx \overline{W}_{L,\lambda}(t)(1 + \frac{\hat{\lambda}' t\overline{W}_{L,\lambda}(t)}{N(t)}) \end{aligned} \tag{3.42}$$

3.8.6 An approximating quadratic arrival rate function

If the arrival rate function is not approximately linear, then a quadratic approximation can be considered:

$$\lambda(s) \approx \lambda_q(s) = a + bs + cs^2 \qquad 0 \leq s \leq t. \tag{3.43}$$

where a, b, and c are constant. Eick et al. [48] studied the quadratic arrival rate functions in an IS system and obtained a nice formula for expected number of busy servers in the system at time t which is mentioned in the following theorem.

Theorem 3.8.2. *Suppose λ is quadratic as in (3.43). If*

$$E[S^3] < \infty$$

Then,

$$E[Q(t)] = E[\lambda(t - S_e)]E[S] + cVar(S_e)E[S].$$

Proof: See Theorem 9 in [48]

As mentioned, the fact is that TVLL in Theorem (3.8.1) is connected to infinite-server (IS) queuing models as the number of customers in the system is equal to the number of busy servers in IS models and the waiting times are equal to service times in IS models. By using this fact and Theorem 3.8.2, $E[L(t)]$ is obtained:

$$E[L(t)] = E[\lambda(t - W_e)]E[W] + cVar(W_e)E[W]. \tag{3.44}$$

Now consider (3.43) to approximate arrival rate and plug moment formula in the equation (3.32) into (3.44) to get:

$$\begin{aligned} E\left[L(t)\right] \approx E\left[L_q(t)\right] &\equiv (a + bs + cs^2)E\left[W\right] - (b + 2ct)\gamma_W^2 E\left[W\right]^2 + 2c\theta_W^3(t)E\left[W\right]^3 \\ &= \lambda_q(t)E\left[W\right] - \gamma_W^2\lambda_q^{'}(t)E\left[W\right]^2 + \theta_W^3(t)\lambda_q^{''}(t)E\left[W\right]^3 \end{aligned} \tag{3.45}$$

where

$$\gamma_W^2 = \frac{(c_W^2 + 1)}{2}, \qquad c_W^2 = \frac{Var(W)}{E\left[W\right]^2}$$

$$\theta_W^3(t) = \frac{E\left[W^3\right]}{6E\left[W\right]^3}$$

$$\lambda_q^{'} = b + 2ct$$

$$\lambda_q^{''}(t) = 2c.$$

By integrating over $[0, t]$ and dividing by t in (3.45), it is obtained:

$$E\left[\overline{L}(t)\right] \equiv t^{-1}\int_0^t E\left[L(s)\right] ds \approx \overline{\lambda}_q(t)E\left[W\right] - \gamma_W^2\overline{\lambda_q^{'}}(t)E\left[W\right]^2 + \theta_W^3(t)\lambda_q^{''}(t)E\left[W\right]^3 \tag{3.46}$$

where

$$\overline{\lambda}_q(t) \equiv t^{-1}\int_0^t \lambda_q(s)ds = a + b(\frac{t}{2}) + c(\frac{t^2}{3}) \tag{3.47}$$

and

$$\overline{\lambda_q^{'}}(t) \equiv t^{-1}\int_0^t \lambda_q^{'}(s)ds = b + ct. \tag{3.48}$$

Plug in $x \equiv E\left[W\right]$ and $\overline{L}(t) \approx E\left[\overline{L}(t)\right]$ into the equation (3.46):

$$\theta_W^3(t)\lambda_q^{''}(t)x^3 - \gamma_W^2\overline{\lambda}_q^{'}(t)x^2 + \overline{\lambda}_q(t)x - \overline{L}(t) = 0 \tag{3.49}$$

To solve this equation of degree 3, perturbation method can be used by assuming

$$\lambda_q^{''} \ll \overline{\lambda_q^{'}}(t) \ll \overline{\lambda}_q(t)$$

$$x(\epsilon) = x_0 + \epsilon x_1 + o(\epsilon^2)$$

$$\lambda_q^{''} = O(\epsilon) \qquad as \qquad \epsilon \to 0$$

and then using (3.38) for the $O(1)$ terms. As a result, we get the following approximation:

$$x(\epsilon) \equiv \overline{W}_{L,\lambda,q}(t) \approx W_{L,\lambda,q,p}(t) \equiv w(1 + w\delta + w^2\epsilon(\frac{1}{1 - 2w\delta})) \tag{3.50}$$

where

$$w \equiv \overline{W}_{L,\lambda}(t) \equiv \frac{\overline{L}(t)}{\overline{\lambda}(t)}$$

$$\delta \equiv \frac{\gamma_W^2 \overline{\lambda}_q^{'}(t)}{\overline{\lambda}_q(t)}$$

$$\epsilon \equiv \frac{\theta_W^3(t)\lambda_q^{''}}{\overline{\lambda}_q(t)}$$

with $\epsilon \ll \delta \ll 1$.

3.8.7 An approximating cubic arrival rate

If the arrival rate function is neither approximately linear nor approximately quadratic, then a polynomial function of degree 3 can be considered to approximate arrival rate function:

$$\lambda(s) \approx \lambda_c(s) = a + bs + cs^2 + es^3, \qquad a \neq 0, \qquad 0 \leq s \leq t \tag{3.51}$$

where a, b, c and e are constants. To connect TVLL to the infinite-server (IS) when arrival rates are approximated by a cubic function, we refer to Theorem 3.8.3.

Theorem 3.8.3. *Consider a $M_t\backslash G\backslash\infty$ queue. Let $\lambda^{(k)}$ denote the kth derivative of λ. For any $n \geq 0$, suppose λ is $n+1$-times differentiable and $(n+1)^{st}$ derivative is Riemann integrable on $[t-x,t]$ for all x. If*

$$E[S^{n+2}] < \infty$$

and

$$E[\lambda^{(k)}(t - S_e^{k+1})] < \infty, \qquad 0 \le k \le n+1$$

then,

$$E[Q(t)] = m_n(t) + R_n(t) \tag{3.52}$$

where

$$m_n(t) = \sum_{k=1}^{n} (-1)^k \frac{\lambda^{(k)}(t) E[S^{k+1}]}{(k+1)!}$$

$$R_n(t) = (-1)^{n+1} E[\lambda^{(n+1)}(t - S_e^{n+2})] \frac{E[S^{n+2}]}{(n+2)!}.$$

Proof: See Theorem 10 in [48].

Given the fact that the $L(t)$ is equal to $Q(t)$ in IS models and the waiting times W are equal to service times S in IS models and considering (3.51) to approximate arrival rate, Theorem 3.8.3 can be applied to get the following formula:

$$E[L(t)] \approx E[L_c(t)]$$

$$= \lambda_c(t) E[W] - \gamma_W^2 \lambda_c^{'}(t) E[W]^2 + \theta_W^3(t) \lambda_c^{''}(t) E[W]^3 - \alpha_W^4(t) \lambda_c^{'''}(t) E[W]^4 \tag{3.53}$$

where

$$\gamma_W^2 = \frac{(c_W^2 + 1)}{2}, \qquad c_W^2 = \frac{Var(W)}{E[W]^2}$$

$$\theta_W^3(t) = \frac{E\left[W^3\right]}{6E[W]^3}$$

$$\alpha_W^4(t) = \frac{E\left[W^4\right]}{24E[W]^4}$$

$$\lambda_c^{'} = b + 2cs + 3es^2$$

$$\lambda_c^{''} = 2c + 6es$$

$$\lambda_c^{'''} = 6e.$$

By integrating over $[0, t]$ and dividing by t in (3.53), it is obtained:

$$E\left[\overline{L}(t)\right] \equiv t^{-1} \int_0^t E\left[L(s)\right] ds$$

$$\approx \overline{\lambda}_c(t) E\left[W\right] - \gamma_W^2 \overline{\lambda_c^{'}}(t) E\left[W\right]^2 + \theta_W^3(t) \overline{\lambda_c^{''}}(t) E\left[W\right]^3 - \alpha_W^4(t) \lambda_c^{'''} E\left[W\right]^4 \tag{3.54}$$

where

$$\overline{\lambda}_c(t) = t^{-1} \int_0^t \lambda_3(s) ds = a + b(\frac{t}{2}) + c(\frac{t^2}{3}) + e(\frac{t^3}{4}) \tag{3.55}$$

and

$$\overline{\lambda_c^{'}}(t) = t^{-1} \int_0^t \lambda_q^{'}(s) ds = b + ct + et^2 \tag{3.56}$$

$$\overline{\lambda_c^{''}}(t) = t^{-1} \int_0^t \lambda_q^{''}(s) ds = 2c + 3et. \tag{3.57}$$

Consider $x \equiv E\left[W\right]$ and $\overline{L}(t) \approx E\left[\overline{L}(t)\right]$, then substitute in the equation (3.54):

$$\alpha_W^4(t) \lambda_c^{'''}(t) x^4 - \theta_W^3(t) \overline{\lambda_c^{''}}(t) x^3 + \gamma_W^2 \overline{\lambda_c^{'}}(t) x^2 - \overline{\lambda}_c(t) x + \overline{L}(t) = 0 \tag{3.58}$$

Now, the root of the equation (3.58) should be estimated. To solve the equation, perturbation method can be applied by assuming

$$\lambda_c^{'''} \ll \overline{\lambda_c^{''}} \ll \overline{\lambda_c^{'}}(t) \ll \overline{\lambda}_c(t)$$

$$x(\epsilon) = x_0 + \epsilon x_1 + o(\epsilon^2)$$

$$\lambda_q^{'''} = O(\epsilon) \qquad as \qquad \epsilon \to 0$$

First, let's write the equation (3.58) in a simple format. The equation (3.58) can be divided by $\overline{\lambda}_c(t)$ (Since $a \neq 0$, then $\overline{\lambda}_c(t) \neq 0$)

$$\frac{\alpha_W^4(t)\lambda_c^{'''}(t)}{\overline{\lambda_c}(t)}x^4 - \frac{\theta_W^3(t)\overline{\lambda_c^{''}}(t)}{\overline{\lambda_c}(t)}x^3 + \frac{\gamma_W^2\overline{\lambda}_c^{'}(t)}{\overline{\lambda_c}(t)}x^2 - \frac{\overline{\lambda}_c(t)}{\overline{\lambda}_c(t)}x + \frac{\overline{L}(t)}{\overline{\lambda}_c(t)} = 0 \tag{3.59}$$

If we define

$$\epsilon = \frac{\alpha_W^4(t)\lambda_c^{(3)}(t)}{\overline{\lambda}_c(t)}$$

$$\beta = \frac{\theta_W^3(t)\overline{\lambda}_c^{''}(t)}{\overline{\lambda}_c(t)}$$

$$\delta = \frac{\gamma_W^2\overline{\lambda}_c^{'}(t)}{\overline{\lambda}_c(t)}$$

$$\omega = \frac{\overline{L}(t)}{\overline{\lambda}_c(t)}$$

and plug in them into (3.59), we get:

$$\epsilon x^4 + \beta x^3 + \delta x^2 + x - \omega = 0 \tag{3.60}$$

where

$$\epsilon \ll \beta \ll \delta \ll 1.$$

Substitute $x(\epsilon) = a_0 + a_1\epsilon$ into the equation (3.60) to get:

$$\epsilon(a_0 + a_1\epsilon)^4 + \beta(a_0 + a_1\epsilon)^3 + \delta(a_0 + a_1\epsilon)^2 + (a_0 + a_1\epsilon) - \omega = 0 \tag{3.61}$$

Then collect the powers of ϵ and use Theorem 3.6.1 which leads to the following equations:

$$\beta a_0^3 + \delta a_0^2 + a_0 - \omega = 0 \tag{3.62}$$

$$a_0^4 + 3\beta a_1 a_0^2 - 2\delta a_0 a_1 + a_1 = 0 \tag{3.63}$$

Again, perturbation theory is used to obtain a_0 in the equation (3.62)

$$a_0 = \omega(1 + \delta\omega - \beta\omega^2(\frac{1}{1 - 2\omega\delta}))$$

Then, a_1 can be obtained easily from equation (3.63)

$$a_1 = \frac{a_0^4}{3\beta a_0^2 - 2\delta a_0 + 1} \approx \frac{\omega^4}{3\beta\omega^2 - 2\delta\omega + 1}$$

As a result, a new estimate of waiting time is obtained in the form of:

$$
\begin{aligned}
W_{L,\lambda,c,p}(t) &\equiv x(\epsilon) = a_0 + a_1\epsilon + O(\epsilon^2) \\
&\approx \omega(1 + \delta\omega - \frac{\beta\omega^2}{1 - 2\omega\delta} + \frac{\epsilon\omega^3}{1 - 2\delta\omega + 3\beta\omega^2})
\end{aligned}
\tag{3.64}
$$

4. Service time in monopolistic walk-in clinics

In this chapter, first we define four scenarios based on walk-in clinics' capacity (finite or infinite) and their position in the area (monopolistic or oligopoly). Then, we focus on clinics acting in a monopolistic market to obtain the optimum value for service time considering patients' satisfaction and clinics' revenue.

4.1 Introduction

In healthcare, walk-in clinics, also called "rapid access clinics" or "medical clinics", refer to any healthcare center providing care without an appointment. In some cases they are owned by doctors who work there, and in some cases the clinics are owned by a larger business which owns multiple clinics and provides physicians with physical and administrative infrastructure.

Walk-in clinics work on a fee-for-service model, so they benefit from the number of patients they serve. Moreover, doctors are paid by health insurance company on a fee-for-service basis and direct a percentage of their payments to the clinic, which employs reception and nursing staff as it considers necessary.

As the number of patients increases, more revenue is gained. Hence, it may be in interest of some walk-in clinics to reduce their service times to increase profit.

On the other hand, short service time sacrifices the quality of service and leads to the dissatisfaction of patients. Patients want to be heard carefully and be asked directly why they have come to the clinic. This is essential especially with patients with multiple medical problems. Adequate service time to communicate with patients could positively influence health outcomes by increasing patient satisfaction, leading to greater patient understanding of health problems and treatments available, contributing to better adherence to treatment plans. In contrast, limiting a patient's complaints during a visit may result in missing important information. A study in USA found that just 36% of doctors posed an open-ended question to get patients to talk and after a doctor asks a question, patients get a median time of 11 seconds to answer before the doctor interrupts them. Sometimes, the doctor interrupts to get clarity from a patient, but 11 seconds is still too soon.[1]

As mentioned, the problem is that clinics tend to allocate less service time than expected by patients. The problem gets worse in rush hours when the number of arrivals has increased but the number of servers could not be increased due to limitation in the number of doctors. In the previous section, we reviewed a well-managed system in which the waiting times often remain approximately stationary, even though the arrival rate is time-varying. That was primarily achieved by using appropriate time-varying staffing. With appropriate staffing, customers do not have to wait in the line and the time spent in the system is almost equal to the service times. However, in the reality in walk-in clinics, we usually have limited number of doctors. When all available doctors are busy, patients have to wait in waiting room. Long service time can lead to overcrowding in clinics with infinite capacity. This issue is also more highlighted in clinics with finite capacity because a certain number of patients are admitted. When waiting room is full, an arriving patient will be turned away. Rejection causes dissatisfaction of refused patients and losing revenue that would be gained by them.

In this section, we study walk-in clinics in which arrival rates are changing over time intervals, while the number of servers can not be changed as much as required.

4.2 Models and analyses

To define the scenarios, the difference between monopolistic and oligopoly market should be clarified. First, definition of these markets is mentioned:

- The Monopolistic Market: A monopolistic market is a theoretical condition describing a market in which there is only one company offering products and services to the public. A monopolistic market is the opposite of a perfectly competitive market. For instance, electricity is an example of a monopoly market. Generally, it is controlled or monitored by the

[1] https://www.cbc.ca/news/health/doctor-patient-visits-1.4755498

governments to safeguard the customers interests.

- The Oligopoly Market: The term oligopoly has been derived from two Greek words: ‘oligi’ which means few and ‘polein’ that means to sell. Oligopoly is a market structure in which there are two or more firms selling products or providing services to customers. Oligopoly is also known as ‘competition among the few’ as there are few sellers in the market and every seller has a great effect and is affected by the behaviour of other firms. For example, the banking industry in Canada is dominated by six big banks: National Bank of Canada, Royal Bank of Canada (RBC), the Bank of Montreal, Canadian Imperial Bank of Commerce, the Bank of Nova Scotia (Scotiabank), and Toronto Dominion Bank (TD).

Therefore, a monopoly is when there is a single walk-in clinic, while an oligopoly is when there are a small number of walk-in clinics in an area. Based on clinics’ capacity (finite or infinite) and position of clinics (monopolistic or oligopoly), four scenarios are defined:

- **Scenario 1 (Model UM)**: The capacity of waiting room is infinite which means no patient is rejected and all arriving patients are served sooner or later. Furthermore, there is no walk-in clinic in the region except this one. The model associated with this scenario is named Model UM, where U represents the uncapacitated system and M shows the monopolistic position of the clinic in the area.

- **Scenario 2 (Model CM)**: The waiting room of clinic has finite capacity and just a limited number of patients can be admitted. When there is no capacity in the waiting room, a new arriving patient will be turned away. Moreover, we are still on the monopolistic position. The model of this scenario is called Model CM showing a capacitated monopolistic clinic.

- **Scenario 3 (Model UO)**: This scenario considers a clinic with infinite capacity in an oligopoly market. In other words, there are several walk-in clinics in the region and all arriving patients to the considered clinic are admitted. The model associated with this scenario is called Model UO, where U indicates an uncapacitated system, and O denotes the oligopoly situation.

- **Scenario 4 (Model CO)**: In this scenario, we consider a clinic in an area in which there are some other walk in clinics. Also, the clinic admits certain number of patients because of limited space. The model associated with this scenario is called Model CO that is suitable to describe many real-life situations.

For each scenario, a mathematical model will be presented . The notations used in this section is presented in Table 4.1.

Table 4.1: Models notations

Parameter	Definition
$\lambda(t)$	Time-varying arrival rate
μ	Service rate
s	Average service time
π_i	Probability that there are i patients in the system
k	Maximum number of available doctors
θ	Patients sensitivity to service time
γ	Patients sensitivity to difference between ideal and allocated service time
β	Level of dissatisfaction caused by overcrowding
s_m	Minimum service time that should be allocated to patients
s_W	Ideal service time
c	Clinic's capacity
α	Patients dissatisfaction caused by being refused
P_i	Patients satisfaction in scenario $i \in \{UM, CM, UO, CO\}$
$s^*_{P_i}$	service time maximizing satisfaction function in scenario i
R	Revenue function
B	Maximum budget assigned by the governments
C_1	Ancillary cost function
C_2	Main cost function
R_N	New revenue function
N	Net profit function

4.2.1 Assumptions

1. Customers arrive according to a Poisson process with rate $\lambda(t)$.

2. Time-between-arrivals follow exponential distributed with mean $\frac{1}{\lambda(t)}$.

3. Service times are exponentially distributed with mean μ.
4. The maximum number of available doctors is k.
5. Queue discipline is first come, first served (FCFS).
6. A doctor serves only one customer at a time.
7. When a patient joins the queue (is admitted to the clinic), s/he will never leave until being visited by a doctor.

4.2.2 Ideal service time s_W and minimum service time s_m

As defied in Table 4.1, in all models the notations s_W and s_m will be used. Before discussing different models, these notations should be explained in more details.

service time is recognized as one of the important factors in patient satisfaction. As the service time increases, satisfaction is enhanced. In a well-organized clinic, ideal service time would be allocated to patients and there is no rush even in rush hours. This is due to the fact that the number of doctors would be increased as needed. When the arrival rate changes (increase or decrees), the number of doctors changes (increase or decrees) to serve arriving patients in an unhurried manner.

However, in reality it is not always possible to provide appropriate time-varying doctors. Clinics with limited number of servers could not allocate ideal service time especially when the number of arriving patients is relatively large compared to the available doctors. Patients usually show sensitivity to difference between ideal service time and time that would be allocated to them. As the difference increases, satisfaction decreases.

In the previous section, a well-managed system with time-varying arrival rate was studied to calculate sojourn time. As mentioned, in this system customers do not wait in the queue and receive service upon arrival. Hence, total waiting time is almost equal to service time. We will use $\overline{W}_{L,\lambda}(t)$, $\overline{W}_{L,\lambda,l}(t)$, $\overline{W}_{L,\lambda,l,p}(t)$, $\overline{W}_{L,\lambda,q,p}(t)$, $\overline{W}_{L,\lambda,c,p}(t)$ as the estimation of ideal service time in systems with linear, quadratic and cubic arrival rate.

Although it could not be always expected to provide ideal service time and highest level of satisfaction, service time should be as long as patients have positive level of satisfaction. In addition, s_m can be calculated such that satisfaction function in each model is greater than zero:

$$P_i(s_m) \geq 0 \qquad i \in \{UM, CM, UO, CO\}.$$

4.3 Model UM

This model shows a clinic which acts as a monopoly in the region. It has infinite capacity with time-varying arrival rate and k doctors. Such clinic with the addressed assumptions is modeled as a $M_t/M/k/\infty$ queuing system.

Since all arriving patients are accepted, people may experience overcrowding in such clinic when the number of patients being present in the clinic exceeds the number of servers. In addition, overcrowding in healthcare center is defined as having more patients than staff who should ideally care for. It is also known as dangerously crowded because delays in services may have unpleasant consequences which go well beyond the inconvenience of spending hours in the line. It may lead to an increased risk of medical errors, delayed access to treatments, and increased gridlock in the broader health care system. Hence, arriving at a crowded walk-in clinic increases the risk of bad experience and dissatisfaction.

In this model, satisfaction of patients is presented in a functional form as shown bellow:

$$P_{um}(s) = \theta s - \gamma(s_W - s) - \beta \Sigma_{i=k+1}^{\infty} \pi_i \tag{4.1}$$

$$s_m \leq s \leq s_W$$

Patient satisfaction increases by longer service time while decrease by overcrowding and difference between ideal service time and allocated time. In the following, calculation of π_i will be discussed.

4.3.1 Calculating π_i

For a multi-server queue, π_i can be calculated by using birth-death process. When a birth occurs, the process goes from state i to $i+1$, while in a death the process goes from state i to state $i-1$. As obtained in the equation (3.9), the process is specified by arrival rates $\{\lambda_i\}_{i=1,2,..}$ and service rates $\{\mu_i\}_{i=1,2,...}$.

$$\pi_i = \frac{\lambda_0\lambda_1\lambda_2...\lambda_{i-1}}{\mu_1\mu_2\mu_3...\mu_i}\pi_0. \tag{4.2}$$

Figure 4.1 indicates state diagram for a $M_t/M/k/\infty$ model. As shown, the service rate for states smaller than k is equal to $i\mu$, while for states larger than k is equal to $k\mu$:

$$\mu_i = \mu = \begin{cases} i\mu & i < k \\ k\mu & i \geq k \end{cases} \tag{4.3}$$

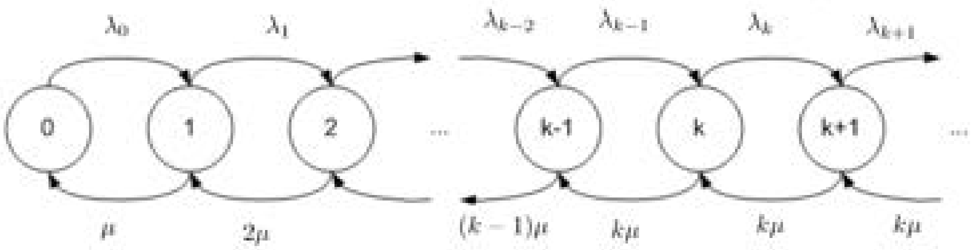

Figure 4.1: State diagram for a multi-server model

Based on considered estimation for arrival rates (linear, quadratic and cubic) in the interval $[0, T]$, λ_i is defined:

$$\lambda_i = \overline{\lambda} = \begin{cases} \overline{\lambda}_l(T) & \text{if } \textit{linear} \\ \overline{\lambda}_q(T) & \text{if } \textit{quadratic} \\ \overline{\lambda}_c(T) & \text{if } \textit{cubic} \end{cases} \tag{4.4}$$

where

$$\overline{\lambda}_l(T) = \frac{1}{T}\int_0^T (\hat{a}_0 + \hat{a}_1 t)dt = \hat{a}_0 + \hat{a}_1(\frac{T}{2}), \tag{4.5}$$

$$\overline{\lambda}_q(T) = \frac{1}{T}\int_0^T (\hat{a}_0 + \hat{a}_1 t + \hat{a}_2 t^2)dt = \hat{a}_0 + \hat{a}_1(\frac{T}{2}) + \hat{a}_2(\frac{T^2}{3}), \tag{4.6}$$

$$\overline{\lambda}_c(T) = \frac{1}{T}\int_0^T (\hat{a}_0 + \hat{a}_1 t + \hat{a}_2 t^2 + \hat{a}_3 t^3)dt = \hat{a}_0 + \hat{a}_1(\frac{T}{2}) + \hat{a}_2(\frac{T^2}{3}) + \hat{a}_3(\frac{T^3}{4}). \tag{4.7}$$

Therefore, π_i can be calculated as follows:

$$\pi_i = \frac{\lambda_0}{\mu_1} \times \frac{\lambda_1}{\mu_2} \times \ldots \times \frac{\lambda_{i-1}}{\mu_i} \times \pi_0 \approx \frac{\overline{\lambda}}{\mu} \times \frac{\overline{\lambda}}{2\mu} \times \ldots \times \frac{\overline{\lambda}}{i\mu} \times \pi_0 = \frac{\overline{\lambda}^i}{i!\mu^i}\pi_0, \quad i \le k \tag{4.8}$$

$$\pi_i = \frac{\lambda_0}{\mu_1} \times \frac{\lambda_1}{\mu_2} \times \ldots \times \frac{\lambda_{i-1}}{\mu_i} \times \pi_0 \approx \frac{\overline{\lambda}^k}{k!\mu^k}\left(\frac{\overline{\lambda}}{k\mu}\right)^{i-k}\pi_0, \qquad i \ge k \tag{4.9}$$

To find π_0, we use the fact that the sum of all transition probabilities equals 1. Thus:

$$1 = \sum_{j=0}^{\infty}\pi_j = \sum_{j=0}^{k-1}\frac{\overline{\lambda}^j}{j!\mu^j}\pi_0 + \sum_{j=k}^{\infty}\frac{\overline{\lambda}^k}{k!\mu^k}\left(\frac{\overline{\lambda}}{k\mu}\right)^{j-k}\pi_0 \tag{4.10}$$

In other words:

$$\pi_0 = \frac{1}{\sum_{j=0}^{k-1}\frac{\overline{\lambda}^j}{j!\mu^j} + \sum_{j=k}^{\infty}\frac{\overline{\lambda}^k}{k!\mu^k}\left(\frac{\overline{\lambda}}{k\mu}\right)^{j-k}}. \tag{4.11}$$

Based on the equation (3.1)

$$s = E[t_{service}] = \frac{1}{\mu}.$$

Substitute μ with $\frac{1}{s}$ in the equation (4.11) to get:

$$\pi_0 = \frac{1}{\sum_{i=0}^{k-1} \frac{(\bar{\lambda} s)^i}{i!} + \sum_{i=k}^{\infty} \frac{(\bar{\lambda} s)^k}{k!} \left(\frac{\bar{\lambda} s}{k} \right)^{i-k}}. \tag{4.12}$$

Consequently,

$$\pi_i = \frac{(\bar{\lambda} s)^k}{k!} \left(\frac{\bar{\lambda} s}{k} \right)^{i-k} \frac{1}{\sum_{j=0}^{k-1} \frac{(\bar{\lambda} s)^j}{j!} + \sum_{j=k}^{\infty} \frac{(\bar{\lambda} s)^k}{k!} \left(\frac{\bar{\lambda} s}{k} \right)^{j-k}} \qquad i \geq k+1 \tag{4.13}$$

4.3.2 Concavity property of P_{um}

Concavity property can be used to obtain the global maximum of P_{um}. Proposition 4.3.1 shows P_{um} has this property.

Proposition 4.3.1. *P_{um} is a concave function.*

Proof: Let us consider arbitrary points $x_1 < x_2$ in the interval $[0, T]$. Assume

$$x_0 = tx_1 + (1-t)x_2, \quad t \in [0, T]$$

By using the mean value version of Taylor's theorem, it is obtained:

$$P_{um}(x_1) = P_{um}(x_0) + P_{um}^{'}(x_0)(x_1 - x_0) + \frac{1}{2} P_{um}^{''}(\xi_1)(x_1 - x_0)^2$$

$$P_{um}(x_2) = P_{um}(x_0) + P_{um}^{'}(x_0)(x_2 - x_0) + \frac{1}{2} P_{um}^{''}(\xi_2)(x_2 - x_0)^2$$

where

$$x_1 \leq \xi_1 \leq x_0$$

$$x_0 \leq \xi_2 \leq x_2.$$

Since $P^{''}_{um} = -\beta\Sigma^{\infty}_{i=k+1}\pi^{''}_{i} \leq 0$, it can be concluded:

$$P_{um}(x_1) \leq P_{um}(x_0) + P^{'}_{um}(x_0)(x_1 - x_0),$$

$$P_{um}(x_2) \leq P_{um}(x_0) + P^{'}_{um}(x_0)(x_2 - x_0).$$

The result of multiplying the $P_{um}(x_1)$ by t and $P_{um}(x_2)$ by $(1 - t)$ and adding is:

$$tP_{um}(x_1) + (1 - t)P_{um}(x_2) \leq P_{um}(x_0) + P^{'}_{um}(x_0)(tx_1 + (1 - t)x_2 - x_0) = P_{um}(tx_1 + (1 - t)x_2).$$

As a result, P_{um} is a concave function.

Proposition 4.3.2. *Any local maximum of function P_{um} is also a global maximum.*

Proof: Let x^* be a local maximum of P_{um}. So, there is a $\delta > 0$ such that for $x \in (x^* - \delta, x^* + \delta)$:

$$P_{um}(x) \leq P_{um}(x^*).$$

Suppose towards a contradiction that there exists $\hat{x}$ such that

$$P_{um}(\hat{x}) > P_{um}(x^*).$$

Consider the line segment

$$x(t) = t\hat{x} + (1 - t)x^*, \quad t \in [0, 1].$$

Due to concavity property of P_{um},

$$P_{um}(x(t)) \geq P_{um}(x^*), \qquad t \in [0,1]$$

since

$$P_{um}(t\hat{x} + (1-t)x^*) \geq tP_{um}(\hat{x}) + (1-t)P_{um}(x^*) > tP_{um}(x^*) + (1-t)P_{um}(x^*) = P_{um}(x^*).$$

Now, t can be picked sufficiently close to 0 such that

$$x(t) \in (x^* - \delta, x^* + \delta).$$

Therefore

$$P_{um}(x(t)) \leq P_{um}(x^*)$$

by the definition of $(x^* - \delta, x^* + \delta)$. This is a contradiction. Hence, it follows that

$$P_{um}(x) \leq P_{um}(x^*)$$

for all $x \in [0, T]$. Therefore, x^* is a global maximum of P_{um}.

4.3.3 Optimization P_{um}

We seek to solve optimization problem

$$\max_{s} P_{um}.$$

Due to concavity property of P_{um}, it is enough to find its local maximum. To find the local maximum, the root of $P^{'}_{um}$ should be found. Obtaining a closed-form solution is not possible, so the numerical methods will be applied. In this research, we focus on Newton–Raphson method which is also known as the Newton method. In calculus, it is an iterative method for finding the root of a differentiable function.

The first step in this method is putting initial guess of the root which is typically denoted by s_0 with the true root represented by s^*_{um}. Therefore, the true root can be represented as:

$$s^*_{um} = s_0 + t,$$

where t shows how far the guess is from the true value of the root. As t is small, a linear tangent line is used to approximate the location of the root which is written as:

$$0 = P^{'}_{um}(s^*_{um}) = P^{'}_{um}(s_0 + t) \approx P^{'}_{um}(s_0) + tP^{''}_{um}(s_0)$$

Therefore, t can be estimated:

$$t \approx -\frac{P^{'}_{um}(s_0)}{P''_{um}(s_0)}.$$

Combining this approximation with the true value s^*_{um} yields:

$$s^*_{um} = s_0 + t \approx s_0 - \frac{P^{'}_{um}(s_0)}{P''_{um}(s_0)}.$$

So, the new estimate of s^*_{um}, s_1 is

$$s_1 = s_0 - \frac{P^{'}_{um}(s_0)}{P''_{um}(s_0)}.$$

If we continue, the iteration of Newton-Raphson will be achieved:

$$s_{i+1} = s_i - \frac{P^{'}_{um}(s_i)}{P''_{um}(s_i)}.$$

Generating sequence is stopped when the the difference between two successive ones is less than considered error e:

$$|s_{i+1} - s_i| < e.$$

4.3.4 Revenue

In walk-in clinics, revenue is gained by serving patients. More arriving patients generate more revenue. Revenue in this model does not depend on service time, because regardless of quality of service, all patients in the region choose the clinic. All arriving patients are admitted and served, so the clinic neither benefits nor loses by modifying the average service time. Hence, the clinic can be asked to allocate average service time s^*_{um} to maximize patients' satisfaction.

4.4 Model CM

In this scenario, there is a cap on the number of admitted patients and due to limited capacity, all arriving patients could not be admitted . It is the case for those walk-in clinics that are the only clinic in a region and their waiting room capacity is small relative to the arriving patients. There is a noticeable difference between patient satisfaction in models UM and model CM. Moreover, in model UM the expected satisfaction values are the same for all arriving patients, while in model CM the values for accepted and rejected patients are different. It is assumed that the refused patients have a fixed level of dissatisfaction α. Therefore, to maximize the expected patient satisfaction, the regulator solves:

$$P_{cm}(s) = (1 - \pi_c)(\theta s - \gamma s_W + \gamma s) - \alpha \pi_c \tag{4.14}$$

$$s_m \leq s \leq s_W$$

where π_c is the probability that the system contains c patients and therefore a new arriving patient is turned away. In the following, π_c will be calculated.

4.4.1 Calculating π_c

As obtained in the equation (3.9), for queues with k servers π_c is equal to:

$$\pi_c = \frac{\lambda_0\lambda_1\lambda_2 ... \lambda_{k-1}\lambda_k\lambda_{k+1} ... \lambda_{c-1}}{\mu_1\mu_2 ... \mu_{k-1}\mu_k\mu_{k+1} ... \mu_c}\pi_0. \tag{4.15}$$

Given (4.9), π_c can be calculated as follows:

$$\pi_c = \frac{\overline{\lambda}^k}{k!\mu^k}\left(\frac{\overline{\lambda}}{k\mu}\right)^{c-k}\pi_0. \tag{4.16}$$

By using the fact that the sum of all transition probabilities equals 1, π_0 can be obtained:

$$1 = \sum_{i=0}^{\infty}\pi_i = \sum_{i=0}^{c}\pi_i = \sum_{i=0}^{k-1}\frac{\overline{\lambda}^i}{i!\mu^i}\pi_0 + \sum_{i=k}^{c}\frac{\overline{\lambda}^k}{k!\mu^k}\left(\frac{\overline{\lambda}}{k\mu}\right)^{i-k}\pi_0 \tag{4.17}$$

$$\pi_0 = \frac{1}{\displaystyle\sum_{i=0}^{k-1}\frac{\overline{\lambda}^i}{i!\mu^i} + \sum_{i=k}^{c}\frac{\overline{\lambda}^k}{k!\mu^k}\left(\frac{\overline{\lambda}}{k\mu}\right)^{i-k}} \tag{4.18}$$

Furthermore, given the equation (3.1), μ in the equation (4.18) can be substituted with $\frac{1}{s}$,

$$\pi_0 = \frac{1}{\displaystyle\sum_{i=0}^{k-1}\frac{(\overline{\lambda}s)^i}{i!} + \sum_{i=k}^{c}\frac{(\overline{\lambda}s)^k}{k!}\left(\frac{\overline{\lambda}s}{k}\right)^{i-k}}. \tag{4.19}$$

Using the equations (4.16) and (4.19), π_c is obtained as

$$\pi_c = \frac{(\overline{\lambda} s)^c}{k!k^{c-k}}\left(\frac{1}{\sum_{i=0}^{k-1}\frac{(\overline{\lambda} s)^i}{i!} + \sum_{i=k}^{c}\frac{(\overline{\lambda} s)^k}{k!}\left(\frac{\overline{\lambda} s}{k}\right)^{i-k}}\right). \tag{4.20}$$

Proposition 4.4.1. *π_c is an increasing function of s.*

Proof: For simplicity, put

$$t = \overline{\lambda} s$$

$$h = \frac{1}{k!k^{-k}},$$

and define the function f as

$$f(t) = \frac{t^c}{\sum_{i=0}^{k-1}\frac{t^i}{i!} + h\sum_{i=k}^{c}\frac{t^i}{k^i}}.$$

Therefore, π_c in the equation (4.20) can be written in the form of

$$\pi_c = \frac{f(t)}{hk^c}, \tag{4.21}$$

To prove π_c is an increasing function, it is enough to show $f(t)$ is an increasing function. Let us differentiate f:

$$\frac{\partial f}{\partial t} = \frac{(ct^{c-1})(\sum_{i=0}^{k-1} \frac{t^i}{i!} + h\sum_{i=k}^{c} \frac{t^i}{k^i}) - (t^c)(\sum_{i=0}^{k-1} \frac{it^{i-1}}{i!} + h\sum_{i=k}^{c} \frac{it^{i-1}}{k^i})}{A^2}, \tag{4.22}$$

where

$$A = \sum_{i=0}^{k-1} \frac{t^i}{i!} + h\sum_{i=k}^{c} \frac{t^i}{k^i} = 1 + \sum_{i=1}^{k-1} \frac{t^i}{i!} + h\sum_{i=k}^{c} \frac{t^i}{k^i} > 0.$$

$\frac{\partial f}{\partial t}$ in the equation (4.22) can be written in the form of:

$$\frac{\partial f}{\partial t} = (t^{c-1})\frac{\sum_{i=0}^{k-1} \frac{(c-i)t^i}{i!} + h\sum_{i=k}^{c} \frac{(c-i)t^i}{k^i}}{A^2}. \tag{4.23}$$

Since $c \geq i$, then

$$\frac{\partial f}{\partial t} \geq 0.$$

As a result, f (and consequently π_c) is an increasing function.

4.4.2 Objective function of satisfaction

The goal is to maximize P_{cm} by setting the best value for s. Therefore, the objective function is defined as:

$$\max_s P_{cm} = (1 - \pi_c)(\theta s - \gamma s_W + \gamma s) - \alpha\pi_c \tag{4.24}$$

$$s_m \leq s \leq s_W.$$

To find the optimum service time, the critical points of an n-degree polynomial must be calculated. Based on Abel Ruffini theorem [49] obtaining the closed-form solutions of a general polynomial function with higher than 5 degrees is not analytically possible. We apply numerical method to find the optimum average service time and represent it with s^*_{cm}.

Proposition 4.4.2. *The maximum value of P_{cm} is at $s^*_{cm} > 0$.*

Proof: P_{cm} is is continuously differentiable on $s > 0$ and

$$\lim_{s \to 0} P_{cm} = \frac{\partial P_{cm}}{\partial s} = \theta + \gamma > 0.$$

Therefore, there is an $\epsilon > 0$ such that

$$P_{cm}(s = \epsilon) > P_{cm}(s = 0),$$

concluding that the maximum of P_{cm} is achievable at a value of s greater than zero.

4.4.3 Revenue function

Each patient admitted by the clinic will generate revenue for the walk-in clinic. Hence, revenue is made by serving the accepted patients that is defined as:

$$R(s) = \overline{\lambda}(1 - \pi_c(s)). \tag{4.25}$$

Since, π_c is an increasing function of s, R is a decreasing function of s. In other words, as average service time increases, the probability of rejection in capacitated clinics increases and consequently revenue decreases. To gain more revenue, some clinics tend to reduce service time which affects quality of service and patient care. The main goal of studying model CM is identifying the cases where revenue maximization policies are not aligned with patient care. In the following, it is outlined how this can be handled by the government's plan.

4.4.4 Government's budget

Since revenue function is a decreasing function of s, the maximum revenue would be obtained by allocation of minimum service time s_m. If capacitated walk-in clinics allocate more service time, their revenue will be less than $R(s_m)$

$$s \geq s_m \Rightarrow R(s) \leq R(s_m). \tag{4.26}$$

We would like to know how we can encourage capacitated clinics to increase the quality of service and allocate more than minimum service time to patients. Let us assume

$$b_1 = \frac{R(s)}{R(s_m)} \tag{4.27}$$

and

$$b_2 = 1 - b_1. \tag{4.28}$$

When allocated time is s such that $s \geq s_m$, the clinic loses $b_2 R(s_m)$ amount of revenue. To encourage clinics to increase service time, this loss can be compensated by the government based on the performance of clinics. New revenue can be defined as

$$R_N(s) = R(s) + \frac{s}{s^*_{cm}} b_2 R(s_m). \tag{4.29}$$

Due to the limited budget that would be allocated to this plan, clinics could not increase the service time as much as they want and expect that their all lost revenue could be compensated by the government. In this case, Proposition 4.4.3 and 4.4.4 would be helpful to set a cap for the average service time based on limitation on budget.

Theorem 4.4.1. *Let f be a continuous function defined on $[a, b]$ and let x be a number with $f(a) < x < f(b)$. Then the intermediate value theorem guarantees that there exists some s between a and* *b such that* $f(s) = x$

Proof: See [50]

Proposition 4.4.3. *: Assume B is the maximum budget assigned by the government to this plan and define*

$$d = \frac{B}{R(s_m)}.$$

Then there is a s_d *such that*

$$\pi_c(s_d) = 1 - [(1 - d)(1 - \pi_c(s_m))].$$

Proof: Since $0 \leq d \leq 1$ and $0 \leq \pi_c(s_m) \leq 1$, it can be concluded:

$$0 \leq 1 - [(1 - d)(1 - \pi_c(s_m))] \leq 1. \tag{4.30}$$

Given Theorem 4.4.1 and the fact that $\pi_c(s)$ is a continuous function between 0 and 1, there is s_d such that

$$\pi_c(s_d) = 1 - [(1 - d)(1 - \pi_c(s_m))].$$

Proposition 4.4.4. *Suppose that the average service time allocated to patients is s. Then* $s \leq s_d$ *if only if* $b_2R(s_m) \leq B$.

Proof: Based on Proposition 4.4.1, π_c is an increasing function. Therefore,

$$
\begin{aligned}
s \le s_d &\iff \pi_c(s) \le \pi_c(s_d) \\
&\iff \pi_c(s) \le 1 - [(1-d)(1-\pi_c(s_m))] \\
&\iff (1-d)(1-\pi_c(s_m)) \le (1-\pi_c(s)) \\
&\iff (1-d)\overline{\lambda}(1-\pi_c(s_m)) \le \overline{\lambda}(1-\pi_c(s)) \\
&\iff (1-d)R(s_m) \le R(s) \\
&\iff 1-d \le \frac{R(s)}{R(s_m)} \\
&\iff 1-d \le b_1 \\
&\iff 1-d \le 1-b_2 \\
&\iff b_2 \le d \\
&\iff b_2R(s_m) \le R(s_m)d \\
&\iff b_2R(s_m) \le B.
\end{aligned}
$$

As a consequence,

$$s \leq s_d \Longleftrightarrow \pi_c(s) \leq \pi_c(s_d) \Longleftrightarrow b_2 R(s_m) \leq B. \tag{4.31}$$

4.4.5 Ancillary cost and net profit

Let us look at the new revenue function defined in (4.29),

$$R_N(s) = R(s) + \frac{s}{s^*_{cm}} b_2 R(s_m)$$

or equivalently,

$$R_N(s) = b_1 R(s_m) + \frac{s}{s^*_{cm}} b_2 R(s_m)$$

$$= (1 - b_2) R(s_m) + \frac{s}{s^*_{cm}} b_2 R(s_m).$$

At a glance, it can be seen that

$$R_N(s) \leq R(s_m),$$

because

$$\frac{s}{s^*_{cm}} \leq 1.$$

Now, the question is "what could encourage the clinics to increase the service time, while their total revenue would be less than original revenue $R(s_m)$?" The answer is in ancillary cost and net profit. Ancillary costs are different from main costs. In walk-in clinics it means all costs other than servers employment expenses and cost of renting clinic space. Usually, a percentage of revenue gained from serving patients is assigned to ancillary costs. Let p_1 be this percentage, C_1 be the ancillary cost, C_2 be the main cost, and N be the net profit such that:

$$C_1(s) = p_1 R(s) = p_1[\overline{\lambda}(1 - \pi_c(s))] \tag{4.32}$$

$$N(s) = R_N(s) - C_1 - C_2 = (1 - p_1)R(s) + \frac{s}{s^*_{cm}} b_2 R(s_m) - C_2. \tag{4.33}$$

Proposition 4.4.5. *Assume that*

$$\omega = \frac{s}{s^*_{cm}}.$$

Then $p_1 \geq 1 - \omega$ *if only if* $N(s) \geq N(s_m)$.

Proof: We have

$$p_1 \geq 1 - \omega \Longleftrightarrow (1 - b_1)p_1 \geq (1 - b_1)(1 - \omega)$$

$$\Longleftrightarrow -b_1 p_1 + p_1 \geq 1 - b_1 - \omega + \omega b_1$$

$$\Longleftrightarrow b_1 + \omega - \omega b_1 - b_1 p_1 \geq 1 - p_1$$

$$\Longleftrightarrow (b_1 + \omega - \omega b_1 - b_1 p_1)R(s_m) \geq (1 - p_1)R(s_m)$$

$$\Longleftrightarrow (1 - p_1)b_1 R(s_m) + (1 - b_1)\omega R(s_m) \geq (1 - p_1)R(s_m)$$

$$\Longleftrightarrow (1 - p_1)R(s) + (1 - b_1)\omega R(s_m) \geq (1 - p_1)R(s_m)$$

$$\Longleftrightarrow (1 - p_1)R(s) + (1 - b_1)\omega R(s_m) - C_2 \geq (1 - p_1)R(s_m) - C_2$$

$$\iff (1 - p_1)R(s) + b_2\omega R(s_m) - C_2 \geq (1 - p_1)R(s_m) - C_2$$

$$\iff (1 - p_1)R(s) + b_2\frac{s}{s^*_{cm}}R(s_m) - C_2 \geq (1 - p_1)R(s_m) - C_2$$

$$\iff N(s) \geq N(s_m).$$

Proposition 4.4.5 represents the minimum service time to obtain more net profit. In other words, it emphasizes that to gain more net profit, we should allocate

$$s \geq s^*_{cm}(1 - p_1). \tag{4.34}$$

However, based on the limitation on budget in (4.31), there is a cap on s:

$$s \leq s_d. \tag{4.35}$$

Therefore, clinics first evaluate s_d and $s^*_{cm}(1 - p_1)$ and start to increase service time if

$$s_d \geq s^*_{cm}(1 - p_1). \tag{4.36}$$

Otherwise, their net profit would not be more than when they allocate minimum service time. As a result, to gain more profit, the average service time should be

$$s^*_{cm}(1-p_1) \leq s \leq s_d. \tag{4.37}$$

4.4.6 Objective function of profit

The objective is to maximize the net profit:

$$\max_s N(s) = (1-p_1)R(s) + \frac{s}{s^*_{cm}} b_2 R(s_m) - C_2 \tag{4.38}$$

$$s^*_{cm}(1-p_1) \leq s \leq s_d.$$

To find the optimum value, Proposition 4.4.6 can be applied.

Proposition 4.4.6. *$N(s)$ is an increasing function of s.*

Proof: To determine $N(s)$ is increasing, it is enough to show $\dfrac{\partial N(s)}{\partial s} \geq 0$.

$$N(s) = (1-p_1)R(s) + \frac{s}{s^*_{cm}} b_2 R(s_m) - C_2$$

$$= (1-p_1)R(s) + \frac{s}{s^*_{cm}}[R(s_m) - R(s)] - C_2.$$

Therefore,

$$
\begin{aligned}
\frac{\partial N(s)}{\partial s} &= (1-p_1)\frac{\partial R(s)}{\partial s} + \frac{1}{s^*_{cm}}[R(s_m)-R(s)] - \frac{s}{s^*_{cm}}\frac{\partial R(s)}{\partial s} \\
&= (1-p_1)\frac{\partial R(s)}{\partial s} - \frac{s}{s^*_{cm}}\frac{\partial R(s)}{\partial s} + \frac{1}{s^*_{cm}}[R(s_m)-R(s)] \\
&= -\overline{\lambda}(1-p_1)\frac{\partial \pi_c(s)}{\partial s} + \overline{\lambda}\frac{s}{s^*_{cm}}\frac{\partial \pi_c(s)}{\partial s} + \frac{1}{s^*_{cm}}[R(s_m)-R(s)] \\
&= \overline{\lambda}(-1+p_1+\frac{s}{s^*_{cm}})\frac{\partial \pi_c(s)}{\partial s} + \frac{1}{s^*_{cm}}[R(s_m)-R(s)].
\end{aligned}
\tag{4.39}
$$

Using the Proposition 4.4.1, It can be concluded:

$$\frac{\partial \pi_c(s)}{\partial s} \geq 0. \tag{4.40}$$

Also, since $s^*_{cm}(1-p_1) \leq s$,

$$(-1+p_1+\frac{s}{s^*_{cm}}) \geq 0. \tag{4.41}$$

In addition, the function R defined in the equation (4.25) is a decreasing function of s. Since the average service time allocated to patients must be greater than minimum service time ($s_m \leq s$), then

$$R(s) \leq R(s_m).$$

Therefore,

$$R(s_m) - R(s) \geq 0. \tag{4.42}$$

Given the equation (4.40), (4.41), and (4.42), we conclude that the equation (4.39) is greater than zero:

$$\frac{\partial N(s)}{\partial s} \geq 0.$$

This means that $N(s)$ is an increasing function of s and the value maximizing the function $N(s)$ is s_d. In other words, if the government plans to allocate budget to increase service time and this budget is considerable as clinics can increase service time greater than $s^*_{cm}(1-p_1)$, the clinics should try to use all amount of reward because the maximum net profit is obtained at s_d.

5. Service time in oligopolistic walk-in clinic

In this chapter, first factors contributing to patient's choice of a walk-in clinic will be discussed. Then, Model UO and CO will be review to obtain optimum value for service time considering patient satisfaction and clinic revenue.

5.1 Introduction

The number of walk-in clinics has increased dramatically over the last two decades across the world, with several clinics popping up in one area. So, the structure of health market in such areas is considered to be oligopolistic, with no clinics keeping others from having significant influence. Given the competitive nature of oligopoly market, healthcare organizations make their utmost to attract the maximum number of patients to gain more revenue. It benefits the patients, because competition generally leads to more choice and better quality of service. Decision makers, budget-holders and executives of organizations , who are responsible to identify areas of expenditure and improve profitability, should put patients' needs and preferences at the top of their list of priorities as they carry out their planning. In this section, first the factors contributing to patient's choices of a clinic will be reviewed, then, the optimum value for service time will be obtained in capacitated and uncapacitated walk-in clinics which are providing healthcare services in an oligopoly market.

5.2 Factors contributing towards patient's choice of a walk-in clinic

Walk-in clinics, as the private business in healthcare market, should develop and implement plans for attracting more patients to ensure their survival and success. There are multiple factors contribute to a patient's choice of a healthcare center. Bahadori et al. [51] identified 21 factors that may contribute to patient's choices of a clinic as shown in Table 5.1.

Table 5.1: Factors influencing patient's choice of a clinic

Factors	Variables
Facilities and physical assets	Good facilities and equipment; appropriate clinic environment
Service providers (physicians and employees)	Having good physicians and personnel; being responsive to possible errors; scientific management of the clinic in recent years
Location and place	Having a strategic location
Services	Providing high quality and various services in a day, having all medical disciplines; offering boarding services; using a system for queuing patients properly in all wards
Price	Cheaper free tariffs on visits and para-clinical services; low-cost services for veterans and their families; being a non-commercial clinic; promoting a patient-centered culture
Promotion	The center's reputation; obtaining the top rankings among other centers; direct and indirect advertisements and promotions; the audience of the center

People living in different areas may have different preferences to choose a clinic. For example, people living in an area with high-healthcare costs may prefer a clinic with cheaper tariffs on visits, while people in other region, where all residents receive healthcare at no cost or a very minimal cost, prefer to attend a clinic with high quality of service. Therefore, the most significant contributing factors to attract patients in different areas would be different. In the following, we explain how we can select the best factors based on provided data.

5.2.1 Fitting equation to data

Suppose there is a single dependent variable y and several independent variables $x_1, x_2, ..., x_p$. The purpose is to fit an equation to the data collected on these measurements that explains the dependence of y on $x_1, x_2, ..., x_p$. Equations give very precise and concise descriptions of data explaining how dependent variables are related to independent variables. The equation that generally describes the relationship between y and the independent variables is of the form:

$$y = f(x_1, x_2, ..., x_p | \phi_0, \phi_1, ..., \phi_p) + \epsilon \tag{5.1}$$

where $\phi_0, \phi_1, ..., \phi_p$ are unknown parameters of the function f and ϵ is a random disturbance (usually assumed to have a normal distribution with mean 0 and standard deviation σ).

When fitting models to data, the utmost is to find the simplest form of a model that still adequately describes the relationship between the dependent variable and the independent variables. The linear model as the form of

$$y = \phi_0 + \phi_1 x_1 + \cdots + \phi_p x_p + \epsilon \tag{5.2}$$

is sometimes the first equation to be fitted and only abandoned if it turns out to be inadequate. In many instances, a linear model is the most appropriate model to describe the dependence relationship between the dependent variable and the independent variables. This will be true if the dependent variable increases at a constant rate as any of the independent variables is increased while holding the other independent variables constant. Many non-linear models can be put into the form of a linear model by appropriately transforming the dependent variables and/or any or all of the independent variables. This important fact ensures the wide utility of the linear model (i.e. the fact that many non-linear models can be linearizable). When fitting a multiple linear regression model, independent variables, that are not important in predicting the dependent variable, are likely be included. However, these insignificant variables can be eliminated from the final equation by "finding the best equation strategies". Their purpose is to find the "simplest" model (not containing variables that are not important) yet "adequate" (containing variables that are important). There are several strategies for selecting the best equation:

- Forward selection
- Backward elimination
- Step-wise regression
- All Possible Regressions
- Best Subset Regression.

In the followings, these methods are explained.

Forward selection: This method starts with no variables in the equation. Then, statistical tests are Carries out on variables not in the equation to see which have a significant effect on the dependent variable and the most significant variables are added. The process is continued until all variables not in the equation have no significant effect on the dependent variable.

Backward elimination: This method starts with all variables in the equation. Then, statistical tests are Carries out on variables in the equation to see which have no significant effect on the dependent variable and the least significant variables are deleted. The process is continued until all variables in the equation have a significant effect on the dependent variable.

Step-wise regression: This method uses both forward and backward techniques. It starts with no variables in the equation, Then statistical tests are Carries out on variables not in the equation to see which have a significant effect on the dependent variable and the most significant is added. After adding a variable, it checks to see if any variables added earlier can now be deleted. The process is continued until all variables not in the equation have no significant effect on the dependent variable.

All Possible Regressions: Unlike step-wise, this algorithm tests all possible subsets of the set of potential independent variables. For instance, when there are p independent variables, there will be 2^p subsets of variables and the algorithm fits all regressions involving no regressor, one regressor, two regressors, and so on. Then selection criterion is recorded for each regression. Usually, either adjusted R-squared or Mallows' C_p is the criterion for picking the best fitting models. Once the procedure finishes, the champion for each subset size is determined. We then determine which subset size is optimum for our case. Although this method takes longer time to run than step-wise, it guarantees the right answer. Therefore, when there are 15 or fewer independent variables to choose from, this is the variable selection procedure that should be used.

Best Subset Regression It is similar to all possible regressions. This method can be used when the number of variables is large. In this algorithm, the user supplies the value K and the algorithm identifies the best K subset of $x_1, x_2, ..., x_p$ for predicting y.

All of these methods are procedures for attempting to find the best equation.

5.2.2 Notations and assumptions

In addition to notations used in Table 4.1, new notations will be used in this section which are represented in Table 5.2.

Table 5.2: Model notations

Parameter	Definition
ω_s	Probability of selecting the clinic being studied
ω_i	Probability of selecting clinic i
ω_s^i	Probability that patients rejected from clinic i select clinic being studied
c_i	Capacity of clinic i
π_{c_i}	Probability that there are c_i patients in capacitated clinic i
n	The total number of clinics in an area
n_1	The number of capacitated clinics
n_2	The number of uncapacitated clinics
$\overline{\lambda}_{uo}$	Arrival rate of an uncapacitated walk-in clinic in an oligopoly market
$\overline{\lambda}_{co}$	Arrival rate of an capacitated walk-in clinic in an oligopoly market
s_R	Optimum value maximizing revenue function
C_k	The cost of adding a server
C_c	The cost of adding a capacity

Apart from assumptions mentioned in 4.2.1, we consider two important assumptions in an oligopoly health market:

- The average service time is the only factor contributing to patients' choice of a walk-in clinic.
- There are adequate information (the number of servers, capacity, and service time) of other clinics.

Moreover, while trying to find the best equation, it can be seen that the average service time is the variable which should remain in the equation, because the quality of service is a main determinant of the choice of healthcare providers ([52] and [53]). A study conducted at the Inanda C Community Health Centre (CHC) [54] shows that the most common process indicator that patients agreed on as reasons for attending Inanda C CHC was the average service time. The patients were so satisfied when the doctor or nurse explained their sickness and treatment and they get good quality of care. Therefore, we consider average service time as one of the most important factors contributing towards patient's choice. In this research, it is considered that all clinics are very well located, equipped, advertised with same price and their difference is just in the average service time that they allocate to the patients. Hence, there will be just a variable "s" playing role in patients' choice. Also, in an oligopoly health market, every walk-in clinic is affected by the behavior of other walk-in clinics. Therefore, a clinic acting in an oligopoly market should have enough information about other competitors such as the number of servers, capacity and service time they provide.

5.2.3 The probability of selecting a walk-in clinic by patients

Consider linear relationship between the patient's choice of a walk-in clinic and service time as bellows:

$$y = f(s|\phi) + \epsilon \approx \phi s, \tag{5.3}$$

where y indicates the level of patient happiness which leads to selecting a clinic in a competitive market. When there are n walk-in clinics in the vicinity, the probability of selecting the clinic, shown by ω_s, is defined as:

$$\omega_s = \frac{\phi s}{\phi s_1 + \phi s_2 + \cdots + \phi s_{n-1} + \phi s} = \frac{s}{s_1 + s_2 + \cdots + s_{n-1} + s} \tag{5.4}$$

where $s_1, s_2, ..., s_{n-1}$ demonstrate the average service time in $n-1$ walk-in clinics (it is assumed that there is enough information about the average service time allocated by other clinics in an oligopolistic market). When the number of walk-in clinic in the area increases from 1 to n, the number of patients does not change. Hence, average arrival rate calculated in (4.4) can be used in oligopoly models as well, however the quality of service determines the parentage of patients coming to a clinic. Clinics should monitor healthcare market in a regular basis, because any change in service time of other clinics to absorb more patients will affect the probability of choosing other clinics.

5.3 Model UO

This model shows a clinic which acts in an oligopoly market with unlimited capacity. The difference between Model UO and Model UM is in the arrival rate. In Model UM there is no competitor in the vicinity and patients don't have any other choice. Whether the quality of service is good or bad, they select the clinic. However, in Model UO, patients can compare the quality of service and then select a clinic. Therefore, in Model UO the quality of service is important not only for patients but also for clinics.

5.3.1 Arrival rate

Suppose there are n walk-in clinics in the area such that there exist n_1 capacitated clinics and n_2 uncapacitated clinics:

$$n = n_1 + n_2.$$

Let c_i be the capacity of clinic i and π_{c_i} be the probability of presence of c_i patients in the clinic i. Also, consider ω_i shows the probability of selecting clinic i as defined bellow:

$$\omega_i = \frac{s_i}{s_1 + s_2 + \cdots + s_{n-1} + s}.$$

Only capacitated walk-in clinics reject patients. Consider ω_s^i shows the probability that patients rejected from capacitated clinic i select this uncapacitated walk-in clinic which is defined as:

$$\omega_s^i = \frac{s}{s_1 + s_2 + ... + s_{i-1} + s_{i+1} + ... + s_{n-1} + s}, \quad i = 1, 2, ..., n_1 \tag{5.5}$$

Therefore, the arrival rate for an uncapacitated walk-in clinic serving patients in an oligopoly market is:

$$\overline{\lambda}_{uo} = \omega_s \overline{\lambda} + \sum_{i=1}^{n_1} \omega_s^i \pi_{c_i} \omega_i \overline{\lambda}, \tag{5.6}$$

where $\overline{\lambda}$ is the average arrival rate obtained in (4.4). In this research, we neglect the arrivals rejected from two or more capacitated clinics.

Proposition 5.3.1. *$\overline{\lambda}_{uo}$ is an increasing function of s.*

Proof: Assume that:

$$H = s_1 + s_2 + ... + s_{n-1}.$$

Then $\overline{\lambda}_{uo}$ defiend in equation (5.6) can be written in the form of:

$$\overline{\lambda}_{uo} = \overline{\lambda} \left(\frac{s}{H + s} + \sum_{i=1}^{n_1} \frac{\pi_{c_i} s s_i}{(H - s_i + s)(H + s)} \right).$$

To prove $\overline{\lambda}_{uo}$ is increasing, we show $\dfrac{\partial \overline{\lambda}_{uo}}{\partial s} \geq 0$.

$$\frac{\partial \overline{\lambda}_{uo}}{\partial s} = \overline{\lambda}\left(\frac{H}{(H+s)^2} + \sum_{i=1}^{n_1} \pi_{c_i} s_i \left(\frac{H^2 - Hs_i - s^2}{(H - s_i + s)^2 (H+s)^2}\right)\right)$$

$$= \frac{\overline{\lambda}}{(H+s)^2}\left(H + \sum_{i=1}^{n_1} \pi_{c_i} s_i \left(\frac{H^2 - Hs_i - s^2}{(H - s_i + s)^2}\right)\right)$$

$$= \frac{\overline{\lambda}}{(H+s)^2}\left(H + \sum_{i=1}^{n_1} \pi_{c_i} s_i \left(\frac{H^2 - Hs_i}{(H - s_i + s)^2}\right) - \sum_{i=1}^{n_1} \pi_{c_i} s_i \left(\frac{s^2}{(H - s_i + s)^2}\right)\right)$$

$$\geq \frac{\overline{\lambda}}{(H+s)^2}\left(H + \sum_{i=1}^{n_1} \pi_{c_i} s_i \left(\frac{H^2 - Hs_i}{(H - s_i + s)^2}\right) - \sum_{i=1}^{n_1} s_i\right) \qquad (*)$$

$$= \frac{\overline{\lambda}}{(H+s)^2}\left(\sum_{i=1}^{n_1} \pi_{c_i} s_i \left(\frac{H^2 - Hs_i}{(H - s_i + s)^2}\right) + H - \sum_{i=1}^{n_1} s_i\right).$$

Here inequality in $(*)$ holds since:

$$\pi_{c_i} \leq 1$$

and:

$$\left(\frac{s^2}{(H - s_i + s)^2}\right) \leq 1.$$

Given the fact that:

$$H^2 - Hs_i = H(H - s_i) \geq 0$$

and:

$$H - \sum_{i=1}^{n_1} s_i \geq 0,$$

it is concluded that:

$$\frac{\partial \overline{\lambda}_{uo}}{\partial s} \geq 0.$$

5.3.2 Revenue

In Model UM, the revenue does not depend on quality of service, while in Model UO there is a direct link between revenue and quality of service. In addition, higher service time results in greater patient arrival rate and subsequently more revenue. In this model, the revenue is in the form of:

$$R = \max_s \overline{\lambda}_{uo} = \max_s \left(\omega_s \overline{\lambda} + \sum_{i=1}^{n_1} \omega_s^i \pi_{c_i} \omega_i \overline{\lambda} \right),$$

$$s_m \leq s \leq s_W.$$

Based on Proposition 5.3.1, $\overline{\lambda}_{uo}$ is an increasing function of s. Therefore, maximum revenue is gained at s_W.

5.3.3 Satisfaction

In this model, satisfaction of patients is represented in a functional form as shown below:

$$P_{uo} = \theta s + \gamma(s_W - s) - \beta \sum_{i=k+1}^{\infty} \pi_i \tag{5.7}$$

where:

$$\pi_i = \frac{(\overline{\lambda}_{uo} s)^i}{k!k^{i-k}} \left(\frac{1}{\sum_{j=0}^{k-1} \frac{(\overline{\lambda}_{uo} s)^j}{j!} + \sum_{j=k}^{\infty} \frac{(\overline{\lambda}_{uo} s)^k}{k!} \left(\frac{\overline{\lambda}_{uo} s}{k} \right)^{j-k}} \right). \tag{5.8}$$

The only difference between π_i defined in equation (5.8) and π_i obtained in (4.13) is in arrival rate. In equation (4.13) $\overline{\lambda}$ has been used, while in the equation (5.8) $\overline{\lambda}_{uo}$ is considered as arrival rate. In this model, patient satisfaction depends not only on the average service time but also on the number of servers in the system. Proposition 5.3.2 shows the importance of the number of servers in satisfaction of patients.

Proposition 5.3.2. *In Model UO, the satisfaction of patients in a system with k + j servers is more than the satisfaction of patients in a system with k servers. In other words, if we put:*

$$P_{uo}^{(k)} = \theta s + \gamma(s_W - s) - \beta \sum_{i=k+1}^{\infty} \pi_i^{(k)},$$

where:

$$\pi_i^{(k)} = \frac{t^i}{k!k^{i-k}} \left(\frac{1}{D_k + \frac{t^k}{k!} \frac{k}{k-t}} \right), \tag{5.9}$$

$$t = \overline{\lambda}_{uo} s,$$

$$D_k = \sum_{j=0}^{k-1} \frac{t^j}{j!},$$

then:

$$P_{uo}^{(k)} \leq P_{uo}^{(k+j)} \qquad j = 0, 1, 2, ...$$

Proof: First, let us to calculate $\sum_{i=k+1}^{\infty} \pi_i^{(k)}$:

$$\sum_{i=k+1}^{\infty} \pi_i^{(k)} = \sum_{i=k+1}^{\infty} \frac{t^i}{k!k^{i-k}} \left(\frac{1}{D_k + \frac{t^k}{k!}\frac{k}{k-t}} \right)$$

$$= \frac{t^{k+1}}{k!(k-t)} \left(\frac{1}{D_k + \frac{t^k}{k!}\frac{k}{k-t}} \right)$$

$$= \frac{t^{k+1}}{k!} \left(\frac{1}{(k-t)D_k + k\frac{t^k}{k!}} \right).$$

Similarly, $\sum_{i=k+2}^{\infty} \pi_i^{(k+1)}$ can be obtained as below:

$$\sum_{i=k+2}^{\infty} \pi_i^{(k+1)} = \frac{t^{k+2}}{(k+1)!(k+1-t)} \left(\frac{1}{D_k + \frac{t^k}{k!}(\frac{k+1}{k+1-t})} \right)$$

$$= \frac{t^{k+2}}{(k+1)!} \left(\frac{1}{(k+1-t)\,D_k + (k+1)\,\frac{t^k}{k!}} \right).$$

Given (3.2),

$$\frac{t}{k} \leq 1.$$

Therefore, the followings are concluded:

$$\frac{t}{k} \leq 1 \Longrightarrow \frac{t}{k+1} \leq 1$$

$$\Longleftrightarrow \frac{t}{k+1}\left(\frac{t^{k+1}}{k!}\right) \leq \left(\frac{t^{k+1}}{k!}\right)$$

$$\Longleftrightarrow \frac{t^{k+2}}{(k+1)!} \leq \frac{t^{k+1}}{k!}$$

$$\Longleftrightarrow \frac{t^{k+2}}{(k+1)!}\left(\frac{1}{(k+1-t)D_k + (k+1)\frac{t^k}{k!}}\right) \leq \frac{t^{k+1}}{k!}\left(\frac{1}{(k-t)D_k + k\frac{t^k}{k!}}\right)$$

$$\Longleftrightarrow \sum_{i=k+2}^{\infty} \pi_i^{(k+1)} \leq \sum_{i=k+1}^{\infty} \pi_i^{(k)}.$$

Iterating will result in:

$$\sum_{i=k+1+j}^{\infty} \pi_i^{(k+j)} \leq \sum_{i=k+1}^{\infty} \pi_i^{(k)} \qquad j = 1, 2, ...$$

As a consequence:

$$P_{uo}^{(k)} \leq P_{uo}^{(k+j)} \qquad j = 0, 1, 2, ...$$

5.4 Model CO

In this scenario, there is a threshold for the maximum number of admitted patients because the capacity of waiting room is finite. This scenario is the case for many real-life situations in more populated areas where there are some other walk-in clinics, but there is a cap on the admitted patients due to limited space in this clinic.

5.4.1 Arrival rate

In this Model, the arrival rate is defined as below:

$$\overline{\lambda}_{co} = \omega_s \overline{\lambda} + \sum_{i=1}^{n_1 - 1} \omega_s^i \pi_{c_i} \omega_i \overline{\lambda}, \tag{5.10}$$

where $\overline{\lambda}$ is defined in (4.4).

Proposition 5.4.1. *$\overline{\lambda}_{co}$ is an increasing function of s*

Proof: We omit the proof because it is very similar to prove given for Proposition 5.3.1.

5.4.2 Revenue

In this scenario, there is a cap on the possible number of patients in the waiting room, and having higher average service time means higher arrival rate. However, higher average service time leads to rejection of more patients. In general, the revenue function is in the form of:

$$R(s) = \max_s \left(\overline{\lambda}_{co} (1 - \pi_c^{(k)}) \right) \tag{5.11}$$

where

$$\pi_c^{(k)} = \frac{(\overline{\lambda}_{co} s)^c}{k! k^{c-k}} \left(\frac{1}{\displaystyle\sum_{i=0}^{k-1} \frac{(\overline{\lambda}_{co} s)^i}{i!} + \sum_{i=k}^{c} \frac{(\overline{\lambda}_{co} s)^k}{k!} \left(\frac{\overline{\lambda}_{co} s}{k} \right)^{i-k}} \right), \tag{5.12}$$

as obtained in equation (4.20); however, in this equation $\overline{\lambda}_{co}$ is used as arrival rate.

Proposition 5.4.2. $1 - \pi_c^{(k)}$ *is a decreasing function of s.*

Proof: Given Proposition 4.4.1,

$$\pi_c = \frac{(\overline{\lambda} s)^c}{k! k^{c-k}} \left(\frac{1}{\displaystyle\sum_{i=0}^{k-1} \frac{(\overline{\lambda} s)^i}{i!} + \sum_{i=k}^{c} \frac{(\overline{\lambda} s)^k}{k!} \left(\frac{\overline{\lambda} s}{k} \right)^{i-k}} \right)$$

is an increasing function. Also, based on Proposition 5.4.1, $\overline{\lambda}_{co}$ is increasing. Therefore,

$$\pi_c^{(k)} = \frac{(\overline{\lambda}_{co} s)^c}{k! k^{c-k}} \left(\frac{1}{\displaystyle\sum_{i=0}^{k-1} \frac{(\overline{\lambda}_{co} s)^i}{i!} + \sum_{i=k}^{c} \frac{(\overline{\lambda}_{co} s)^k}{k!} \left(\frac{\overline{\lambda}_{co} s}{k} \right)^{i-k}} \right)$$

is increasing function of s. Hence, $1 - \pi_c^{(k)}$ is a decreasing function of s.

Revenue in this scenario is the multiplication of an increasing function "$\overline{\lambda}_{co}$" and a decreasing function "$1 - \pi_c^{(k)}$". Obtaining the closed-form solution is not possible, so numerical methods could be used to get s_R representing the optimum value maximizing revenue function.

5.4.3 Adding servers and expanding capacity

In a capaciated clinic, revenue can be increased by expanding capacity and employing new staff, because $\pi_c^{(k)}$ which has a negative impact on revenue will decrease. On the other hand, expanding capacity and adding new servers to the system would be costly. Let x be the number of new servers which can lead to serving more hx patients. Consider the cost of adding a server and a capacity is shown with C_k and C_c, respectively. Capacitated walk-in clinics can increase the number of servers and capacity such that difference between new revenue and initial revenue is higher than cost of employing new servers and expanding capacity. In other words:

$$R(s) - R(s_R) - xC_k - hxC_c \geq 0 \Longrightarrow x \leq \frac{R(s) - R(s_R)}{C_k + hC_c}$$

$$\Longrightarrow x \leq \frac{\overline{\lambda} - R(s_R)}{C_k + hC_c}.$$

Since x is an integer,

$$x = 0, 1, 2, ..., \left[\frac{\overline{\lambda} - R(s_R)}{C_k + nC_c}\right],$$

where $[y]$ is the least integer greater than or equal to y.

The goal is to maximize revenue by setting the best value for s, adding servers and expanding capacity. Therefore. the objective function of profit is defined as:

$$N(s, x) = \max_{s,x} R(s, k + x, c + hx) - xC_k - hxC_c \tag{5.13}$$

$$s_m \leq s \leq s_W$$

$$x = 0, 1, 2, ..., \left[\frac{\overline{\lambda} - R(s_R)}{C_k + nC_c}\right].$$

In other words, there are $\left[\frac{\overline{\lambda} - R(s_R)}{C_k + nC_c}\right] + 1$ optimization problems and:

$$\max_{s}\{N(s, 0), N(s, 1), ..., N(s, \left[\frac{\overline{\lambda} - R(s_R)}{C_k - nC_c}\right])\} \tag{5.14}$$

should be obtained. To find the optimum average service time in each problem , the critical points of an polynomial must be calculated. Since obtaining the closed-form solutions of a general polynomial function with higher than 5 degrees is not analytically possible, numerical methods can be applied to find the optimum average service time maximizing profit function.

Proposition 5.4.3. *In capacitated walk-in clinics, the probability of rejection decreases when the number of servers and capacity increase. In other words,*

$$\pi_c^{(k)} \geq \pi_{c+hx}^{(k+x)}$$

Proof: First, let us consider $x = 1$ and show:

$$\pi_c^{(k)} \geq \pi_{c+h}^{(k+1)}.$$

For simplicity, put:

$$t = \overline{\lambda}_{co} s$$

$$D_k = \sum_{i=0}^{k-1} \frac{t^i}{i!}$$

$$Q = \sum_{i=c+1}^{c+h} \frac{t^{(k+1)}}{(k+1)!} \left(\frac{t}{k+1}\right)^{i-(k+1)}.$$

Therefore, $\pi_c^{(k)}$ defined in equation (5.12) can be written in the form of:

$$\pi_c^{(k)} = \frac{t^c}{k!k^{c-k}} \left(\frac{1}{D_k + \sum_{i=k}^{c} \frac{t^k}{k!} \left(\frac{t}{k}\right)^{i-k}} \right),$$

Consequently:

$$\pi_{c+h}^{(k+1)} = \frac{t^{c+h}}{(k+1)!(k+1)^{c+h-(k+1)}} \left(\frac{1}{\sum_{i=0}^{k} \frac{t^i}{i!} + \sum_{i=k+1}^{c+h} \frac{t^{(k+1)}}{(k+1)!} \left(\frac{t}{k+1}\right)^{i-(k+1)}} \right)$$

$$= \frac{t^c}{k!(k+1)^{c-k}} \frac{t^h}{(k+1)^h} \left(\frac{1}{D_k + \frac{t^k}{k!} + \sum_{i=k+1}^{c} \frac{t^{(k+1)}}{(k+1)!} \left(\frac{t}{k+1}\right)^{i-(k+1)} + Q} \right).$$

Comparing definition of $\pi_c^{(k)}$ and $\pi_{c+h}^{(k+1)}$, it is enough to show that:

$$\frac{(k+1)^{c-k+h} \left(D_k + \frac{t^k}{k!} + \sum_{i=k+1}^{c} \frac{t^{(k+1)}}{(k+1)!} \left(\frac{t}{k+1}\right)^{i-(k+1)} + Q \right)}{t^h} \geq k^{c-k} \left(D_k + \sum_{i=k}^{c} \frac{t^k}{k!} \left(\frac{t}{k}\right)^{i-k} \right).$$

Given inequation (3.2),

$$\frac{\overline{\lambda}_{co} s}{k} \leq 1,$$

or equivalently:

$$\frac{t}{k} \leq 1.$$

So, the followings are obtained:

$$
\begin{aligned}
\frac{t}{k} \leq 1 &\Longrightarrow k \geq t \\
&\Longrightarrow k + 1 \geq t \\
&\Longrightarrow (k+1)^h \geq t^h \\
&\Longrightarrow (k+1)^h (k+1)^{c-i} \geq t^h k^{c-i} \\
&\Longrightarrow (k+1)^h \frac{(k+1)^{c-k}}{(k+1)^{i-k}} \geq t^h \frac{k^{c-k}}{k^{i-k}} \\
&\Longrightarrow (k+1)^h t^{i-k} \frac{(k+1)^{c-k}}{(k+1)^{i-k}} \geq t^h t^{i-k} \frac{k^{c-k}}{k^{i-k}} \\
&\Longrightarrow (k+1)^h (k+1)^{c-k} \frac{t^{i-k}}{(k+1)^{i-k}} \geq t^h k^{c-k} \frac{t^{i-k}}{k^{i-k}} \\
&\Longrightarrow (k+1)^h (k+1)^{c-k} \sum_{i=k}^{c} \frac{t^{i-k}}{(k+1)^{i-k}} \geq t^h k^{c-k} \sum_{i=k}^{c} \frac{t^{i-k}}{k^{i-k}} \\
&\Longrightarrow (k+1)^h (k+1)^{c-k} \frac{t^k}{k!} \sum_{i=k}^{c} \frac{t^{i-k}}{(k+1)^{i-k}} \geq t^h k^{c-k} \frac{t^k}{k!} \sum_{i=k}^{c} \frac{t^{i-k}}{k^{i-k}}.
\end{aligned}
\tag{5.15}
$$

Since:

$$(k+1)^h (k+1)^{c-k} \geq t^h k^{c-k},$$

then:

$$(k+1)^h (k+1)^{c-k} \sum_{i=0}^{k-1} \frac{t^i}{i!} \geq t^h k^{c-k} \sum_{i=0}^{k-1} \frac{t^i}{i!}. \tag{5.16}$$

Given inequation (5.15) and (5.16), the followings are concluded:

$$(k+1)^h (k+1)^{c-k} \left(D + \frac{t^k}{k!} \sum_{i=k}^{c} \left(\frac{t}{k+1} \right)^{i-k} \right) \geq t^h k^{c-k} \left(D + \frac{t^k}{k!} \sum_{i=k}^{c} \left(\frac{t}{k} \right)^{i-k} \right)$$

$$\Downarrow$$

$$\frac{(k+1)^{c-k+h} \left(D + \frac{t^k}{k!} + \sum_{i=k+1}^{c} \frac{t^{(k+1)}}{(k+1)!} \left(\frac{t}{k+1} \right)^{i-(k+1)} \right)}{t^h} \geq k^{c-k} \left(D + \sum_{i=k}^{c} \frac{t^k}{k!} \left(\frac{t}{k} \right)^{i-k} \right)$$

$$\Downarrow$$

$$\frac{(k+1)^{c-k+h} \left(D + \frac{t^k}{k!} + \sum_{i=k+1}^{c} \frac{t^{(k+1)}}{(k+1)!} \left(\frac{t}{k+1} \right)^{i-(k+1)} + Q \right)}{t^h} \geq k^{c-k} \left(D + \sum_{i=k}^{c} \frac{t^k}{k!} \left(\frac{t}{k} \right)^{i-k} \right)$$

As a result:

$$\pi_c^{(k)} \geq \pi_{c+h}^{(k+1)}.$$

Similarly, for all $x = 0, 1, 2, ...$ it can be proven that:

$$\pi_c^{(k)} \geq \pi_{c+xh}^{(k+x)}.$$

5.4.4 Satisfaction

In this model, satisfaction of patients is represented in a functional form as shown bellow:

$$P_{co} = (1 - \pi_c^{(k)})(\theta s - \gamma s_W + \gamma s) - \alpha \pi_c^{(k)}. \tag{5.17}$$

Patients' satisfaction depends on the average service time, the number of servers, and capacity of clinic. In Model CO, as the number of servers and capacity increases the satisfaction increases. Because based on Proposition 5.4.3, when servers and capacity increase, $\pi_c^{(k)}$ which has a negative impact on satisfaction function decreases. So, it is the best strategy for decision makers to employ new servers and expand capacity such that:

$$R(s) - R(s_R) - xC_k - hxC_c \geq 0,$$

and then obtain the optimum value maximizing P_{co} by using numerical methods.

6. Numerical analysis

In this chapter, simulated data (not from the real-world) will be used. First, arrival rate function will be fitted to the arrival data for estimating waiting times in a well-managed system, which is approximately equal to service time. Then, these estimations will be used to obtain the optimum value of service time in different satisfaction and revenue functions defined for monopolistic and oligopolistic models in Chapters 3 and 4. Also, the sensitivity of service time and satisfaction function to different parameter inputs will be measured.

6.1 Estimation of Arrival rates and waiting times

In this section, arrival rate functions and waiting times are approximated with 95% confidence intervals for linear, quadratic and cubic functions.

6.1.1 Approximating the arrival rate

To approximate arrival rates with polynomials, first the coefficients should be obtained. In this research, the case of $M_t/M/k_t$ is considered which shows a multi-server queueing model with a non-homogeneous Poisson arrival process (M_t), exponentially distributed service time (the M), and time-varying staffing level (the k_t). The arrival process and service times are considered to be mutually independent. Using the Ordinary Least Squares (OLS) method developed in Section 3.5.1, the linear, quadratic and cubic arrival rate functions will be fitted to the arrival data for the target interval $[0, 8]$. The estimates with 95% confidence intervals are shown in Table 6.1.

Table 6.1: Estimating arrival rate functions with 95% confidence intervals

Arrival rate	$\hat{a}_0$	$\hat{a}_1$	$\hat{a}_2$	$\hat{a}_3$
Linear	36.1 ± 0.5	2.97 ± 0.12	-	-
Quadratic	53.1 ± 0.5	2.185 ± 0.071	-0.0167 ± 0.015	-
Cubic	40.3 ± 0.05	3.25 ± 0.064	-0.017 ± 0.005	0.015 ± 0.006

Since the coefficients of the arrival rate function are estimated, average arrival rate defined in (4.4) can be calculated as bellows:

$$\lambda_i = \overline{\lambda} = \begin{cases} \overline{\lambda}_l(T) & \text{if } \textit{linear} \\ \overline{\lambda}_q(T) & \text{if } \textit{quadratic} \\ \overline{\lambda}_c(T) & \text{if } \textit{cubic} \end{cases} \tag{6.1}$$

where

$$\begin{aligned} \overline{\lambda}_l(T) &= \frac{1}{T}\int_0^T (\hat{a}_0 + \hat{a}_1 t)dt = \hat{a}_0 + \hat{a}_1(\frac{T}{2}) \\ &= \frac{1}{8}\int_0^8 (36.1 + 2.97t)dt = 36.1 + 2.97(\frac{8}{2}) = 48.9 \end{aligned}$$

$$\begin{aligned} \overline{\lambda}_q(T) &= \frac{1}{T}\int_0^T (\hat{a}_0 + \hat{a}_1 t + \hat{a}_2 t^2)dt = \hat{a}_0 + \hat{a}_1(\frac{T}{2}) + \hat{a}_2(\frac{T^2}{3}) \\ &= \frac{1}{8}\int_0^8 (53.1 + 2.185t - 0.0167t^2)dt \\ &= 53.1 + 2.185(\frac{8}{2}) - 0.0167(\frac{8^2}{3}) \approx 61.48 \end{aligned}$$

$$\begin{aligned} \overline{\lambda}_c(T) &= \frac{1}{T}\int_0^T (\hat{a}_0 + \hat{a}_1 t + \hat{a}_2 t^2 + \hat{a}_3 t^3)dt = \hat{a}_0 + \hat{a}_1(\frac{T}{2}) + \hat{a}_2(\frac{T^2}{3}) + \hat{a}_3(\frac{T^3}{4}) \\ &= \frac{1}{8}\int_0^8 (40.3 + 3.25t - 0.017t^2 + 0.015t^3)dt \\ &= 40.3 + 3.25(\frac{8}{2}) - 0.017(\frac{8^2}{3}) + 0.015(\frac{8^3}{4}) \approx 54.86 \end{aligned}$$

6.1.2 Estimating waiting times with TVLL

Table 6.2 shows the approximation of waiting times estimated by different methods for the linear, quadratic and cubic arrival rate functions. The first estimator is the direct and the rest are the indirect estimators based on LL obtained in (3.37), (3.38),(3.50) and (3.64).

Table 6.2: Waiting time estimates with 95% confidence intervals

$\overline{W}_{L,\lambda}(t)$	$\overline{W}_{L,\lambda,l}(t)$	$\overline{W}_{L,\lambda,l,p}(t)$	$\overline{W}_{L,\lambda,q,p}(t)$	$\overline{W}_{L,\lambda,c,p}(t)$
0.983 ± 0.015	1.052 ± 0.017	1.044 ± 0.016	1.045 ± 0.016	1.039 ± 0.015

Waiting times are estimated in an ideal system, where they often remain approximately stationary even though the arrival rate is time-varying. That is primarily achieved by using appropriate time-varying staffing levels. With appropriate staffing, customers do not have to wait in the line and the time spent in the system is almost equal to the service times. Therefore, the waiting times approximated in Table 6.2 are considered as an ideal service time or s_W.

6.2 Models and optimum service time

In the previous section, an ideal system with no limit on the number of servers was reviewed. However, in reality, there exist a limited number of servers. In this section, it is assumed that there are k servers in the system and the model $M_t/M/k$ with limited and unlimited capacity in a monopolistic and oligopolistic market will be studied.

6.2.1 Model UM

This model shows a clinic which acts as a monopoly in the region. It has infinite capacity with time-varying arrival rate and k doctors. Such a clinic with the addressed assumptions is modeled as an $M_t/M/k/\infty$ queuing system.

In this model, satisfaction presented in (4.1) is in the form of:

$$P_{um} = \theta s - \gamma(s_W - s) - \beta \Sigma_{i=k+1}^{\infty} \pi_i \tag{6.2}$$

$$s_m \leq s \leq s_W.$$

Using numerical methods mentioned in Subsection 4.3.3, s^*_{um} is obtained and shown in Tables 6.3 and 6.4 with considering different arrival rates and different number of servers.

Table 6.3: The optimum value of service time, $\theta = 14, \gamma = 4, \beta = 1, k = 25$

s_m	s_W	$\Sigma_{i=26}^{\infty}\pi_i$	s^*_{um}	P_{um}
0.220	0.983	0.5402671	0.5112048	4.72942
0.234	1.052	0.5402671	0.5112048	4.45342
0.232	1.044	0.5402671	0.5112048	4.48542
0.233	1.045	0.5402535	0.4065993	2.598533
0.231	1.039	0.5401791	0.4556599	3.505641

Table 6.4: The optimum value of service time, $\theta = 14, \gamma = 4, \beta = 1, k = 30$

s_m	s_W	$\Sigma_{i=26}^{\infty}\pi_i$	s^*_{um}	P_{um}
0.220	0.983	0.5164232	0.6134313	6.59334
0.234	1.052	0.5164232	0.6134313	6.31734
0.232	1.044	0.5164232	0.6134313	6.34934
0.233	1.045	0.5164276	0.4879125	4.085997
0.231	1.039	0.5164231	0.546788	5.169761

As it is shown in the tables, the number of servers plays a key role in satisfaction of patients; as the number increases, satisfaction level rises. In the followings, sensitivity to different parameter inputs, including θ, γ, β, will be analyzed and the importance of the number of servers in happiness of patients will be reviewed in more details.

Sensitivity to θ

Obviously, θ values impact the satisfaction of patients in different scenarios. In this section, diffident values for θ will be considered and its influence on optimum service time and satisfaction function will be analyzed.

Table 6.5: Different values of θ,
$\gamma = 4, \beta = 1, k = 40, s_W = 1.052$

θ	s^*_{um}	P_{um}	θ	s^*_{um}	P_{um}
1	0.8179418	-0.5973076	11	0.8179418	7.582111
2	0.8179418	0.2206343	12	0.8179418	8.400053
3	0.8179418	1.038576	13	0.8179418	9.217994
4	0.8179525	1.856518	14	0.8179418	10.03594
5	0.8179418	2.67446	15	0.8179418	10.85388
6	0.8179359	3.492402	16	0.8179418	11.67182
7	0.8179593	4.310343	17	0.8179418	12.48976
8	0.8179600	5.128285	18	0.8179418	13.3077
9	0.8179418	5.946227	19	0.8179418	14.12565
10	0.8179418	6.764169	20	0.8179418	14.94359

Table 6.6: Different values of θ,
$\gamma = 4, \beta = 1, k = 40, s_W = 1.045$

θ	s^*_{um}	P_{um}	θ	s^*_{um}	P_{um}
1	0.6505818	-1.406129	11	0.6505818	5.099689
2	0.6505818	-0.7555471	12	0.6505818	5.750271
3	0.6505818	-0.1049653	13	0.6505818	6.400853
4	0.6505818	0.5456165	14	0.6505818	7.051435
5	0.6505818	1.196198	15	0.6505818	7.702017
6	0.6505818	1.84678	16	0.6505818	8.352599
7	0.6505818	2.497362	17	0.6505818	9.00318
8	0.6505818	3.147944	18	0.6505818	9.653762
9	0.6505818	3.798526	19	0.6505818	10.30434
10	0.6505818	4.449108	20	0.6505818	10.95493

Table 6.7: Different values of θ,
$\gamma = 4, \beta = 1, k = 40, s_W = 1.039$

θ	s^*_{um}	P_{um}	θ	s^*_{um}	P_{um}
1	0.729084	-0.9896066	11	0.729084	6.301234
2	0.729084	-0.2605225	12	0.729084	7.030318
3	0.7290848	0.4685615	13	0.729084	7.759402
4	0.729084	1.197646	14	0.729084	8.488486
5	0.729084	1.92673	15	0.729084	9.21757
6	0.729084	2.655814	16	0.729084	9.946654
7	0.729084	3.384898	17	0.729084	10.67574
8	0.729084	4.113982	18	0.729084	11.40482
9	0.729084	4.843066	19	0.729084	12.13391
10	0.729084	5.57215	20	0.729084	12.86299

Table 6.8: Different values of θ,
$\gamma = 4, \beta = 10, k = 40, s_W = 1.052$

θ	s^*_{um}	P_{um}	θ	s^*_{um}	P_{um}
1	0.5732583	-1.549578	11	0.6469679	4.58302
2	0.5830477	-0.9713347	12	0.6530964	5.233044
3	0.5918763	-0.3838033	13	0.6591967	5.889183
4	0.5999782	0.2121767	14	0.6653424	6.551448
5	0.6075448	0.8159824	15	0.6715653	7.219894
6	0.6147028	1.42714	16	0.6779165	7.894623
7	0.6215415	2.045285	17	0.6844574	8.575794
8	0.6281154	2.670134	18	0.691267	9.263634
9	0.6345257	3.301473	19	0.6984504	9.958462
10	0.6407887	3.939139	20	0.706174	10.66073

In Tables 6.5, 6.6 and 6.7, the relationship between θ and P_{um} for different arrival rates can be seen. In all cases, there is a a direct link between θ and P_{um} which is shown in Figure 6.1 as well. In contrast, as arrival rate increases, the level of satisfaction decreases. This is due the fact that the number of servers is not changed while arrival rate changes. Also, from Tables 6.5, 6.6 and 6.7 it

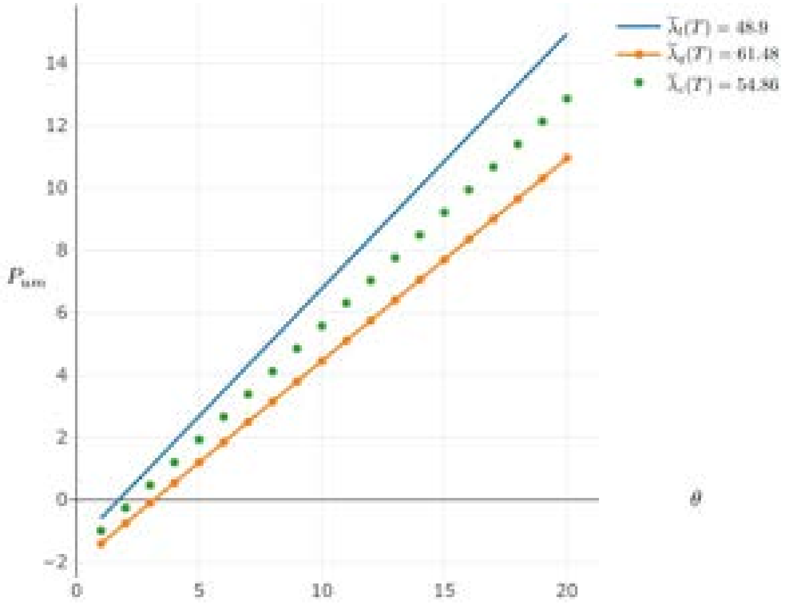

Figure 6.1: Sensitivity of P_{um} to θ

can be seen that θ has no impact on optimum service time s^*_{um}. Precisely, increasing one unit of θ has led to rising satisfaction by s^*_{um}. As mentioned in Subsection 3.2.8, in this research a stable system is studied, where $s \leq \dfrac{k}{\lambda}$. If there is a θ^* such that

$$s^*_{um} = Max\{\frac{k}{\lambda}, s_W\},$$

then for all $\theta \geq \theta^*$ the optimum value of service time remains unchanged, yet satisfaction increases as it is shown in (6.3):

$$\begin{aligned} P^{\theta^*+1}_{um} - P^{\theta^*}_{um} &= ((\theta^* + 1)s^*_{um} - \gamma(s_W - s^*_{um}) - \beta\Sigma^{\infty}_{i=k+1}\pi_i) \\ &- (\theta^* s^*_{um} - \gamma(s_W - s^*_{um}) - \beta\Sigma^{\infty}_{i=k+1}\pi_i) = s^*_{um}. \end{aligned} \tag{6.3}$$

In cases analyzed in Tables 6.5, 6.6 and 6.7, $\theta^* = 1$; when $\theta \geq 1$ the optimum value of service time is not influenced by increasing θ. Let us consider another case to see the effect of θ on s^*_{um}. Table 6.8 demonstrates a particular case with different parameter values where θ impacts both

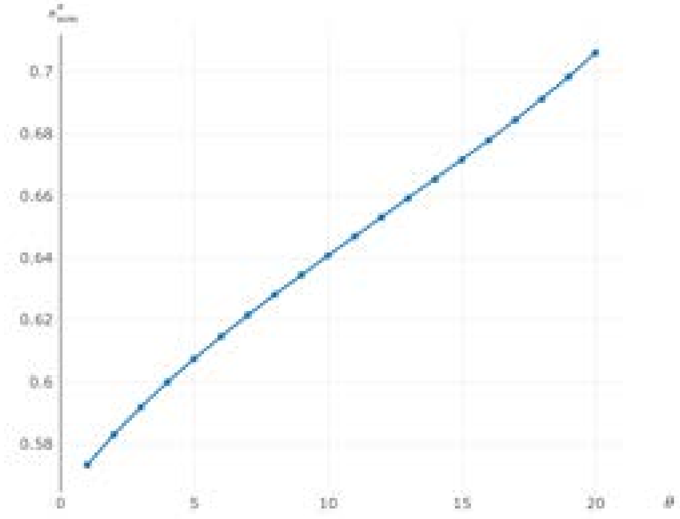

Figure 6.2: Sensitivity of s^*_{um} to θ

satisfaction and the optimum service time. Comparing Table 6.5 and Table 6.8, the only parameter which has been changed is β. This change has affected the results dramatically. As depicted in Table 6.8, both P_{um} and s^*_{um} are rising up as θ increases. Figure 6.2 shows how s^*_{um} is affected by changing θ.

Sensitivity to γ

In this section, the effect of changing γ on optimum service time and satisfaction function will be considered. First, γ is analyzed with considering different values of s_W, then β is perturbed and the influence of γ on optimum service time and satisfaction function is studied. Finally, diffident values for θ and γ will be considered to show the relationship between θ, γ and β.

Table 6.9: Different values of γ,
$\theta = 14, \beta = 1, k = 40, s_W = 1.052$

γ	s^*_{um}	P_{um}	γ	s^*_{um}	P_{um}
1	0.8179418	10.73811	11	0.8179418	8.397529
2	0.8179418	10.50405	12	0.8179418	8.163471
3	0.8179418	10.26999	13	0.8179418	7.929413
4	0.8179418	10.03594	14	0.8179418	7.695355
5	0.8179418	9.801878	15	0.8179418	7.461296
6	0.8179418	9.56782	16	0.8179418	7.227238
7	0.8179418	9.333762	17	0.8179418	6.99318
8	0.8179418	9.099704	18	0.8179418	6.759122
9	0.8179418	8.865645	19	0.8179418	6.525064
10	0.8179418	8.631587	20	0.8179418	6.291006

Table 6.10: Different values of γ,
$\theta = 14, \beta = 1, k = 40, s_W = 1.045$

γ	s^*_{um}	P_{um}	γ	s^*_{um}	P_{um}
1	0.6505818	8.234689	11	0.6505818	4.290508
2	0.6505818	7.840271	12	0.6505818	3.89609
3	0.6505818	7.445853	13	0.6505818	3.501671
4	0.6505818	7.051435	14	0.6505818	3.107253
5	0.6505818	6.657017	15	0.6505818	2.712835
6	0.6505818	6.262599	16	0.6505818	2.318417
7	0.6505818	5.86818	17	0.6505818	1.923999
8	0.6505818	5.473762	18	0.6505818	1.529581
9	0.6505818	5.079344	19	0.6505818	1.135162
10	0.6505818	4.684926	20	0.6505818	0.7407443

Table 6.11: Different values of γ,
$\theta = 14, \beta = 1, k = 40, s_W = 1.039$

γ	s^*_{um}	P_{um}	γ	s^*_{um}	P_{um}
1	0.729084	9.418234	11	0.729084	6.319074
2	0.729084	9.108318	12	0.729084	6.009158
3	0.7290848	8.798402	13	0.729084	5.699242
4	0.729084	8.488486	14	0.729084	5.389326
5	0.729084	8.17857	15	0.729084	5.079411
6	0.729084	7.868654	16	0.729084	4.769495
7	0.729084	7.558738	17	0.729084	4.459579
8	0.729084	7.248822	18	0.729084	4.149663
9	0.729084	6.938906	19	0.729084	3.839747
10	0.729084	6.62899	20	0.729084	3.529831

Table 6.12: Different values of γ,
$\theta = 14, \beta = 10, k = 40, s_W = 1.052$

γ	s^*_{um}	P_{um}	γ	s^*_{um}	P_{um}
1	0.6469679	7.73902	11	0.7146855	4.007073
2	0.6530964	7.337044	12	0.7244131	3.674484
3	0.6591967	6.941183	13	0.736424	3.352638
4	0.6653424	06.551448	14	0.7550057	3.045317
5	0.6715653	6.167894	15	0.817934	2.816531
6	0.6779165	5.790623	16	0.81793775	2.582477
7	0.6844574	5.419794	17	0.81793870	2.348419
8	0.691267	5.055634	18	0.81795890	2.114464
9	0.6984504	4.698462	19	0.81796	1.880429
10	0.706174	4.348727	20	0.81796	1.646389

As presented in Tables 6.9, 6.10 and 6.11, γ has a great impact on P_{um}. An increase in γ causes a decrease in P_{um} which can be seen in Figure 6.3 as well. On the other hand, no change can be seen in the values of s^*_{um}. In addition, when there is a γ^* such that

$$s^*_{um} = Max\{\frac{k}{\lambda}, s_W\},$$

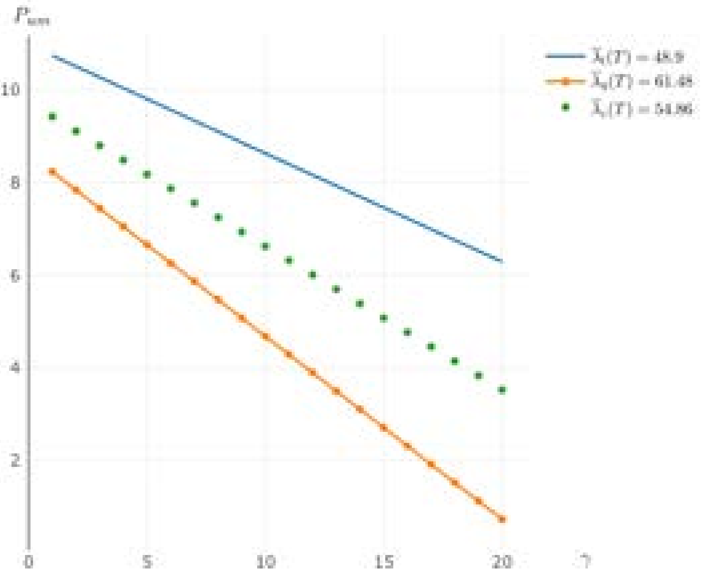

Figure 6.3: Sensitivity of P_{um} to γ

then for all $\gamma \geq \gamma^*$ the optimum value of service time will remain unchanged, while satisfaction will decreased by $s^*_{um} - s_W$, as shown in the following:

$$
\begin{aligned}
& P^{(\gamma^*+1)}_{um} - P^{(\gamma^*)}_{um} \\
&\qquad = (\theta s^*_{um} - (\gamma^* + 1)(s_W - s^*_{um}) - \beta \Sigma^{\infty}_{i=k+1} \pi_i) \\
&\qquad - (\theta s^*_{um} - \gamma^*(s_W - s^*_{um}) - \beta \Sigma^{\infty}_{i=k+1} \pi_i) \\
&\qquad = s^*_{um} - s_W.
\end{aligned}
$$

In cases analyzed in Tables 6.9, 6.10 and 6.11, $\gamma^* = 1$, and for $\gamma \geq 1$, the optimum value of service time is constant. Note that this does not mean that s^*_{um} is not influenced by γ. If the parameter β is changed, then s^*_{um} is affected by γ. In Table 6.12, β has been increased to 10 and consequently s^*_{um} has been influenced. In addition, s^*_{um} is increasing as γ is rising up. However, there is a cap on s^*_{um} and when it reaches out the maximum value, it remains unchanged. Figure 6.4 shows the sensitivity of s^*_{um} to γ.

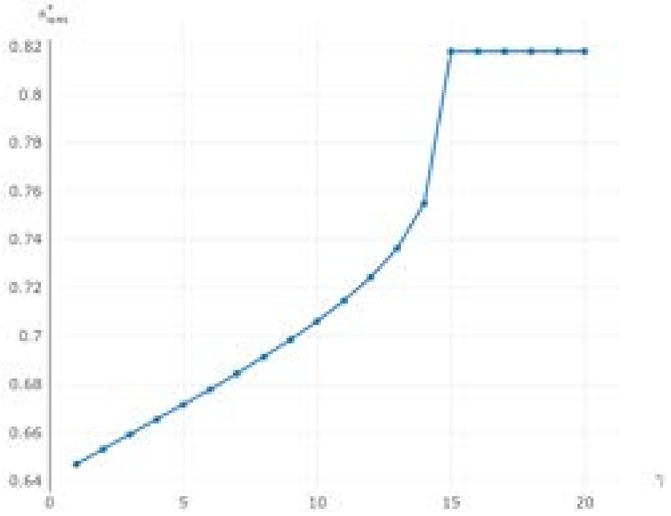

Figure 6.4: Sensitivity of s^*_{um} to γ

It is worth to mention that the impact of θ and γ on s^*_{um} is similar, whereas their effect on P_{um} is different. Furthermore, let us re-write the satisfaction function in (4.1):

$$P_{um} = \theta s - \gamma(s_W - s) - \beta\Sigma_{i=k+1}^{\infty}\pi_i$$

$$= (\theta + \gamma)s - \gamma s_W - \beta\Sigma_{i=k+1}^{\infty}\pi_i.$$

Note that the coefficient of s in this function is $(\theta+\gamma)$. Therefore, both parameters θ and γ have positive effect on s.

Obviously, θ positively impacts P_{um}, while γ does not. In addition, γ is the coefficient of ideal service time s_W, which is always greater than all possible value for s. Therefore:

$$s_W - s^* \geq 0,$$

and consequently, $-\gamma(s_W - s^*)$ always has a negative impact on the satisfaction function.

Relationship between θ, γ and β

In the previous subsections, β was assumed to be equal to 1. It was seen that increasing θ and γ does not affect s^*_{um}. However, If θ and γ are changed simultaneously, then the influence of theses parameters can be seen on s^*_{um}. Table 14 considers different values for θ and γ.

Table 6.13: Different values of θ and γ,
$\beta = 1, k = 40, s_W = 1.052$

θ	γ	s^*_{um}	θ	γ	s^*_{um}
1	0	0.6147028	3	3	0.8179418
0	1	0.6147028	1	4	0.8179418
1	1	0.6779165	2	4	0.8179418
2	1	0.8179352	3	4	0.8179418
3	1	0.8179418	1	5	0.8179418
1	2	0.8179352	2	5	0.8179418
2	2	0.8179418	3	5	0.8179418
3	2	0.8179418	1	6	0.8179418
1	3	0.8179418	2	6	0.8179418
2	3	0.8179418	3	6	0.8179418

From the results depicted in Table 6.13, if:

$$\theta + \gamma \leq 2,$$

then s^*_{um} is affected.

Sensitivity to β

In this section, sensitivity to parameter β is reviewed to examine its effects on s^*_{um} and P_{um}.

Table 6.14: Different values of β,
$\theta = 14, \gamma = 4, k = 40, s_W = 1.052$

β	s^*_{um}	P_{um}	β	s^*_{um}	P_{um}
1	0.8179418	10.00258	11	0.6553124	6.438826
2	0.8179418	9.4902	12	0.6469679	6.341224
3	0.8179418	8.977824	13	0.6398322	6.255219
4	0.8179418	8.465447	14	0.6336209	6.178448
5	0.8179322	7.953026	15	0.6281154	6.109202
6	0.8179377	7.440686	16	0.623217	6.046205
7	0.721456	7.056354	17	0.6187618	5.988478
8	0.6948071	6.84652	18	0.6147028	5.935252
9	0.6779165	6.684361	19	0.6109987	5.885914
10	0.6653424	6.551448	20	0.6075448	5.839965

Table 6.15: Different values of β,
$\theta = 14, \gamma = 4, k = 40, s_W = 1.045$

β	s^*_{um}	P_{um}	β	s^*_{um}	P_{um}
1	0.6505818	7.051435	11	0.5086145	4.126066
2	0.6505818	6.572397	12	0.5027146	4.053666
3	0.6505818	6.093359	13	0.4975982	3.989546
4	0.6505818	5.61432	14	0.4930996	3.93209
5	0.6505767	5.135271	15	0.489087	3.88011
6	0.5743791	4.760808	16	0.4854847	3.832705
7	0.5498103	4.566861	17	0.482182	3.789178
8	0.5349501	4.422053	18	0.4791906	3.748977
9	0.5241348	4.305906	19	0.476427	3.711658
10	0.5156462	4.209023	20	0.4738697	3.676858

As depicted in Tables 6.14, 6.15 and 6.16, in all cases P_{um} is changed by varying β. Figure 6.5 displays how P_{um} is affected by β.

Table 6.16: Different values of β,
$\theta = 14, \gamma = 4, k = 40, s_W = 1.039$

β	s^*_{um}	P_{um}	β	s^*_{um}	P_{um}
1	0.729084	8.488486	11	0.5794765	5.264046
2	0.729084	8.009459	12	0.5721717	5.178161
3	0.729084	7.530432	13	0.5659161	5.102432
4	0.729084	7.051406	14	0.5604484	5.034804
5	0.729084	6.572379	15	0.5556179	4.973789
6	0.7290802	6.093348	16	0.5513055	4.91827
7	0.6350029	5.802684	17	0.547392	4.867389
8	0.6134416	5.621328	18	0.5438319	4.820471
9	0.5990752	5.479638	19	0.5405321	4.776977
10	0.5882057	5.363038	20	0.5375131	4.73647

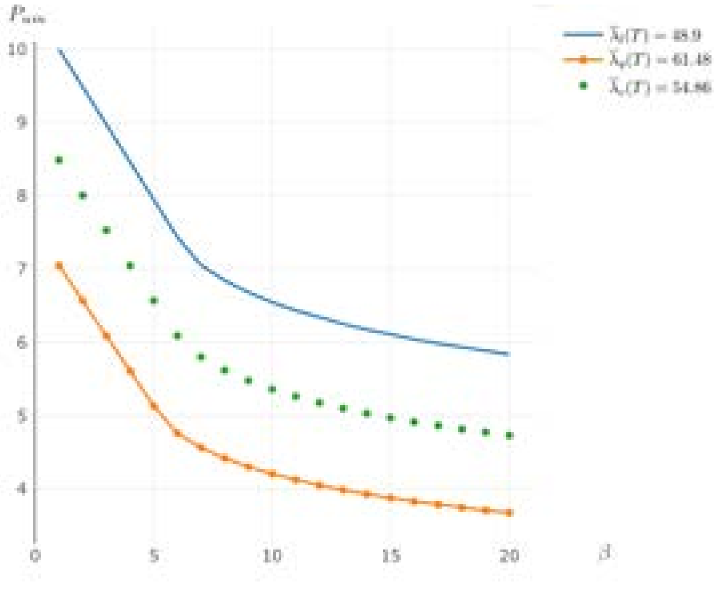

Figure 6.5: Sensitivity of P_{um} to β

As shown in the tables and Figure 6.6, s^*_{um} is changed when β is greater than 5 or 6. In addition, when patients are not so sensitive to overcrowding, servers should concentrate on the patients being served. However, when sensitivity to overcrowding is significant, not only patients who are receiving service but also patients waiting in waiting room should be considered.

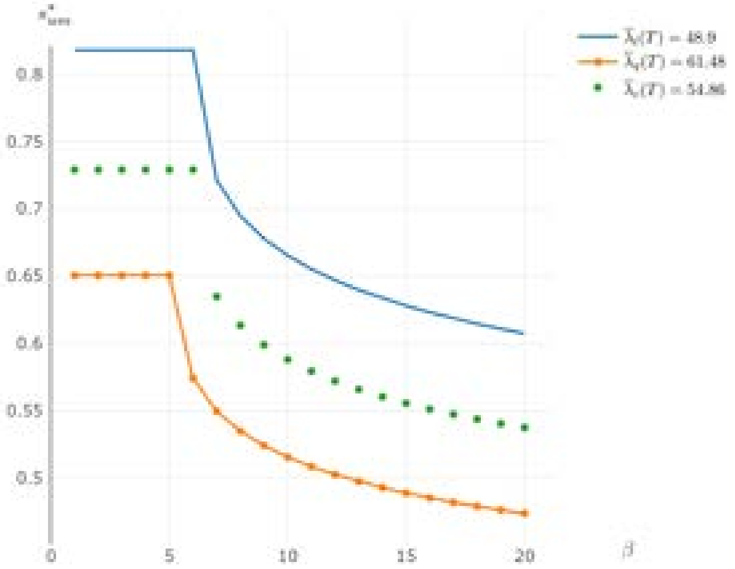

Figure 6.6: Sensitivity of s^*_{um} to β

The importance of *k*

In Table 6.3 and 6.4, the importance of *k* in evaluating satisfaction is visible. In this section, more values for *k* will be considered to show it in more details.

Table 6.17: Different values of k,
$\theta = 14, \gamma = 4, \beta = 1, s_W = 1.052, \overline{\lambda} = 48.9$

k	s^*_{um}	P_{um}	k	s^*_{um}	P_{um}
13	0.2658076	-0.0475548	22	0.4498532	3.332478
14	0.2862517	0.3297666	23	0.4702996	3.706297
15	0.3066959	0.7065033	30	0.6134313	6.31734
16	0.3271417	1.082745	35	0.7157032	8.17826
17	0.3475881	1.458533	40	0.8179418	10.03594
18	0.3680345	1.833907	45	0.9201846	11.89155
19	0.3885142	2.20937	50	1.0224610	13.74601
20	0.4089605	2.584039	51	1.0429090	14.11662
21	0.4294069	2.958399	52	1.0519640	14.30598

Table 6.18: Different values of k,
$\theta = 14, \gamma = 4, \beta = 1, s_W = 1.045, \overline{\lambda} = 61.48$

k	s^*_{um}	P_{um}	k	s^*_{um}	P_{um}
16	0.2602077	-0.0941006	40	0.6505818	7.051435
17	0.2764686	0.2063602	45	0.7319004	8.530426
18	0.2927295	0.5064071	50	0.813219	10.00773
19	0.3089904	0.8060806	55	0.8945407	11.4837
20	0.3252527	1.105434	60	0.9758626	12.95853
21	0.3415154	1.404484	61	0.992127	13.25338
25	0.4065993	2.598533	62	1.008391	13.54819
27	0.4391245	3.194041	63	1.024689	13.84349
30	0.4879125	4.085997	64	1.040953	14.13823
35	0.5692299	5.569834	65	1.044964	14.2397

Table 6.19: Different values of k,
$\theta = 14, \gamma = 4, \beta = 1, s_W = 1.039, \overline{\lambda} = 54.86$

k	s^*_{um}	P_{um}	k	s^*_{um}	P_{um}
14	0.2551569	-0.1779609	25	0.4556599	3.505641
15	0.27338	0.1587976	30	0.546788	5.169761
16	0.2916031	0.4950387	35	0.6379527	6.83074
17	0.3098263	0.8308185	40	0.729084	8.488486
18	0.3280511	1.166208	45	0.8202156	10.14411
19	0.3462762	1.501227	50	0.9113506	11.7981
20	0.3645013	1.835908	54	0.9842587	13.12028
21	0.3827263	2.170281	55	1.002486	13.45071
22	0.4009847	2.504829	56	1.020746	13.78161
23	0.4192098	2.838666	57	1.038965	14.11182

As shown in Table 6.17, if the arrival rate is equal to $\overline{\lambda} = 48.9$, then at least 14 servers are required to allocate minimum service time s_m. In this case, if the managers of walk-in clinics employ less than 14 servers, they can not obtain positive satisfaction of patients. Moreover, the table depicts the maximum number of servers. Employing 52 servers will lend to the clinic's highest performance in satisfying patients.

In Table 6.18, a quadratic arrival rate is considered ($\overline{\lambda} = 61.48$). In this case, the minimum and maximum servers required are 17 and 65 respectively.

In Table 6.19, a walk-in clinic with cubic arrival rate is considered ($\overline{\lambda} = 54.86$). In this clinic, at least 15 servers should be present in the clinic to serve patients. Also, to allocate maximum service time, 57 servers are needed.

From Figure 6.7, the sensitivity of s^*_{um} to k can be seen. k has positive effect on s^*_{um}; however, when s^*_{um} reaches out its maximum, it is not influenced by increasing k any more.

Similarly, there is a direct link between P_{um} and k as shown in Figure 6.8. When the number of servers increases, patients become more happy. Also, it shows having less than minimum number of servers in the system could cause the negative level of satisfaction.

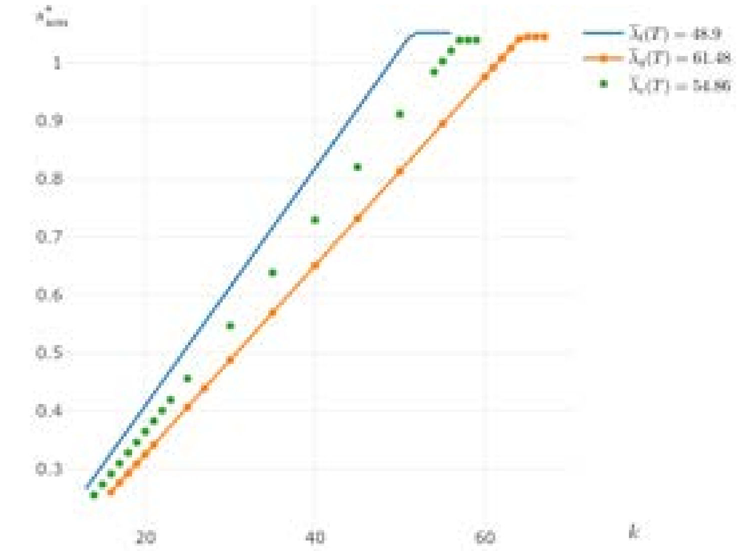

Figure 6.7: Sensitivity of s^*_{um} to k

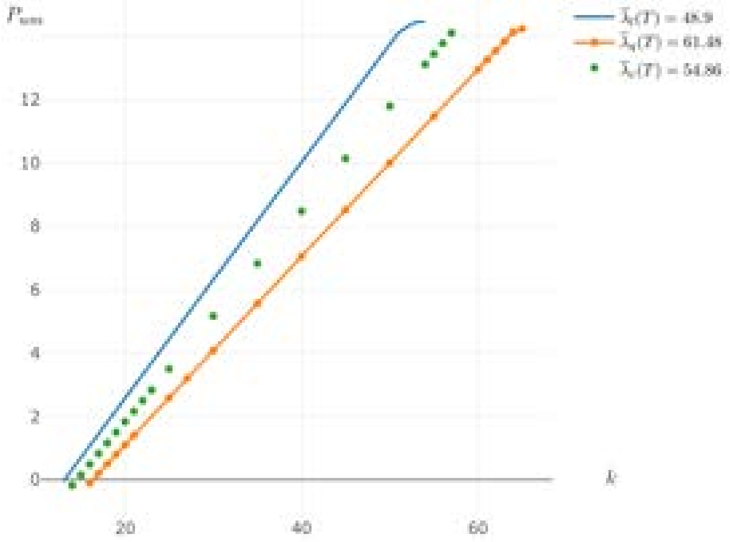

Figure 6.8: Sensitivity of P_{um} to k

Summary of Model UM

The main results obtained for Model UM are summarized in the followings:

- In this model, there is no competitor in the region and all patients are served in this clinic. Thus, the only concern of the clinic is maximizing satisfaction of patients without being worried about revenue.

- In chapter 3, for linear arrival rate two waiting times are obtained ($\overline{W}_{L,\lambda,l}(t)$ and $\overline{W}_{L,\lambda,l,p}(t)$). No matter which one is used as s_W, there is no significant difference between the results.

- As arrival rate increases, less service time can be allocated to patients.

- Impact of different parameters is shown in Table 6.20:

Table 6.20: Impact of different parameters in Model UM

Parameters	s^*_{um}	P_{um}
k	↗	↗
θ	↗	↗
γ	↗	↘
β	↘	↘

- The number of servers plays a key role in satisfaction of patients; as the number increases, satisfaction level rises.

- As sensitivity to service time increases, more service time should be allocated to patients for capturing their maximum satisfaction.

- There is usually difference between ideal service time that patients expected and service time that can be allocated to patients regarding the sources available in the clinic. When patients are so sensitive to this difference, clinics should pay attention to this sensitivity and allocate more service time even if it leads to dissatisfaction of other patients for waiting in the line longer.

- Patients who are receiving service should be always the top priority for walk-in clinics. When patients are not so sensitive to overcrowding, servers should concentrate on the patients being served and spend their expected service time. However, when patients are highly sensitive to overcrowding, the time of visit should be decreased slightly.

6.2.2 Model CM

In this scenario, there is a cap on the number of admitted patients due to limited capacity. Based on 4.14, satisfaction function is as shown below:

$$P_{cm} = (1 - \pi_c)(\theta s - \gamma s_W + \gamma s) - \alpha \pi_c.$$

Let us assume $\theta = 14, \gamma = 4, \alpha = 4$ to obtain π_c, s^*_{cm} and P_{cm}. The results are represented in Table 6.21.

Table 6.21: Optimum value of service time,
$\theta = 14, \gamma = 4, \alpha = 1, k = 25, c = 50$

s_m	s_W	π_c	s^*_{cm}	P_{cm}
0.23377	1.052	0.03124379	0.5111928	4.806211
0.23205	1.044	0.03124379	0.5111928	4.837211
0.23222	1.045	0.03125154	0.4065993	3.009443
0.23089	1.039	0.03124719	0.4556599	3.888209

The interesting result is that system tries to keep π_c stable, while arrival rates are different. This is achieved by differing s^*_{cm}. Also, different levels of satisfaction for different arrival rates can be seen, because there are 25 servers with 50 capacity in the system. As the arrival rate increases, theses numbers of servers and capacity can not satisfy the patients properly.

In the following sections, s^*_{cm} and P_{cm} will be evaluated with different arrival rates and parameters.

Linear arrival rate

In this section, arrival rate is considered linear $\overline{\lambda} = 48.9$ and sensitivity to different parameters is evaluated. First, different values for k are considered.

Table 6.22: Different values of k,
$\theta = 14, \gamma = 4, \alpha = 1, c = 50$

k	s^*_{cm}	P_{cm}	k	s^*_{cm}	P_{cm}
11	0.2248859	-0.2125149	20	0.4089605	2.99566
12	0.2453634	0.1379492	22	0.4498532	3.718536
13	0.2658076	0.4904089	24	0.4907459	4.443269
14	0.2862517	0.8447173	25	0.5111928	4.806211
15	0.3066959	1.200547	30	0.6134313	6.625326
16	0.3271417	1.557655	35	0.7157032	8.450195
17	0.3475881	1.915837	40	0.8179418	10.27797
18	0.3680345	2.274933	45	0.9201846	12.10819
19	0.3885142	2.635078	50	1.022461	13.9409

As shown, at least 12 servers are required and less than this number could not satisfy the patients. Also, increasing the numbers of servers will lead to increasing s^*_{cm} and P_{cm}. Let us review the role of capacity in satisfaction as well.

Table 6.23: Different values of c,
$\theta = 14, \gamma = 4, \alpha = 1, k = 25$

c	s^*_{cm}	P_{cm}	c	s^*_{cm}	P_{cm}
25	0.5111928	4.80571	33	0.5111928	4.80587
26	0.5111928	4.80573	37	0.5111928	4.805951
27	0.5111928	4.80575	40	0.5111928	4.806011
28	0.5111928	4.80577	45	0.5111928	4.806111
29	0.5111928	4.80579	50	0.5111928	4.806211
30	0.5111928	4.80581	55	0.5111928	4.806311

Although increasing capacity results in increasing satisfaction, it is not as important as increasing the number of servers which is shown in Figure 6.9. In addition, increasing capacity leads to leveling up satisfaction slightly, while adding servers causes improving satisfaction hugely.

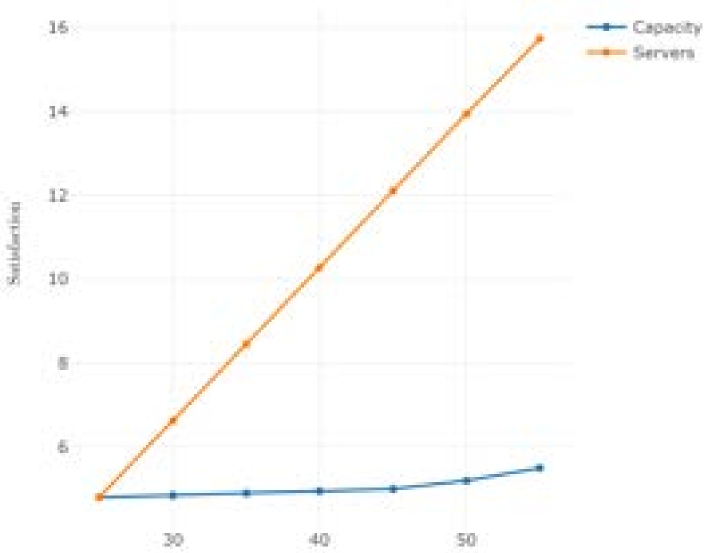

Figure 6.9: Sensitivity of P_{cm} to k and c

In contrast, s^*_{cm} is not affected by c in this case. This is due to the high sensitivity of service time. Comparing with 50 capacity, when capacity is 25, more patients are turned away which can lead to their unhappiness. However, it does not mean that due to their satisfaction, we should decreases service time of the patients in the system. However, when patients are not so sensitive to service time, s^*_{cm} is influenced by c. In Table 6.24, the results depict that increasing one capacity has led to increasing service time by 0.0218052.

Table 6.24: Different values of c,
$\theta = 1, \gamma = 4, \alpha = 14, k = 25$

c	s^*_{cm}
25	0.3655726
26	0.3873778

In the following, sensitivity to the parameters θ, γ and α will be investigated in more details.

First, look at the Table 6.25 in which sensitivity of s^*_{cm} and P_{cm} to θ is presented. As shown, for $k = 25$, $c = 50$ and $\theta \leq 4$, positive satisfaction will never be obtained because other parameters have more impact. To obtain positive satisfaction, k and c should be increased. Also, it is shown that s^*_{cm} is not sensitive to θ, while P_{cm} is. The main reason has been mentioned in Subsection 6.2.1.

Table 6.25: Different values of θ,
$\gamma = 4, \alpha = 1, k = 25, c = 50$

θ	s^*_{cm}	P_{cm}	θ	s^*_{cm}	P_{cm}
1	0.5111928	-1.631664	11	0.5111928	3.320548
2	0.5111928	-1.136443	12	0.5111928	3.815769
3	0.5111928	-0.6412218	13	0.5111928	4.31099
4	0.5111928	-0.1460006	14	0.5111928	4.806211
5	0.5111928	0.3492206	15	0.5111928	5.301432
6	0.5111928	0.8444417	16	0.5111928	5.796654
7	0.5111928	1.339663	17	0.5111928	6.291875
8	0.5111928	1.834884	18	0.5111928	6.787096
9	0.5111928	2.330105	19	0.5111928	7.282317
10	0.5111928	2.825326	20	0.5111928	7.777538

Table 6.26: Different values of γ,
$\theta = 14, \alpha = 1, k = 25, c = 50$

γ	s^*_{um}	P_{cm}	γ	s^*_{cm}	P_{cm}
1	0.5111928	9.349269	11	0.5111928	4.110166
2	0.5111928	8.825359	12	0.5111928	3.586255
3	0.5111928	8.301449	13	0.5111928	3.062345
4	0.5111928	7.777538	14	0.5111928	2.538434
5	0.5111928	7.253628	15	0.5111928	2.014524
6	0.5111928	6.729717	16	0.5111928	1.490614
7	0.5111928	6.205807	17	0.5111928	0.9667031
8	0.5111928	5.681897	18	0.5111928	0.4427927
9	0.5111928	5.157986	19	0.5111928	-0.0811177
10	0.5111928	4.634076	20	0.5111928	-0.605028

Reviewing sensitivity to γ in Table 6.26, it can be seen that when sensitivity to difference between ideal service time and allocated service time is greater than 19 ($\gamma \geq 19$), we need to increase the number of servers and capacity to capture at least the minimum satisfaction of patients. Furthermore, s^*_{cm} and P_{cm} show the different sensitivity to γ that reason has been mentioned in Subsection 6.2.1.

Finally in this section, sensitivity to α is evaluated.

Table 6.27: Different values of α,
$\theta = 14, \gamma = 4, k = 25, c = 50$

α	s^*_{cm}	P_{cm}	α	s^*_{cm}	P_{cm}
1	0.5111928	4.806211	11	0.5111928	4.493774
2	0.5111928	4.774967	12	0.5111928	4.46253
3	0.5111928	4.743724	13	0.5111928	4.431286
4	0.5111928	4.71248	14	0.5092007	4.400732
5	0.5111928	4.681236	15	0.5067591	4.372483
6	0.5111928	4.649992	16	0.5045227	4.346302
7	0.5111928	4.618749	17	0.5024628	4.321917
8	0.5111928	4.587505	18	0.5005869	4.299108
9	0.5111928	4.556261	19	0.4988554	4.277691
10	0.5111928	4.525017	20	0.497248	4.257513

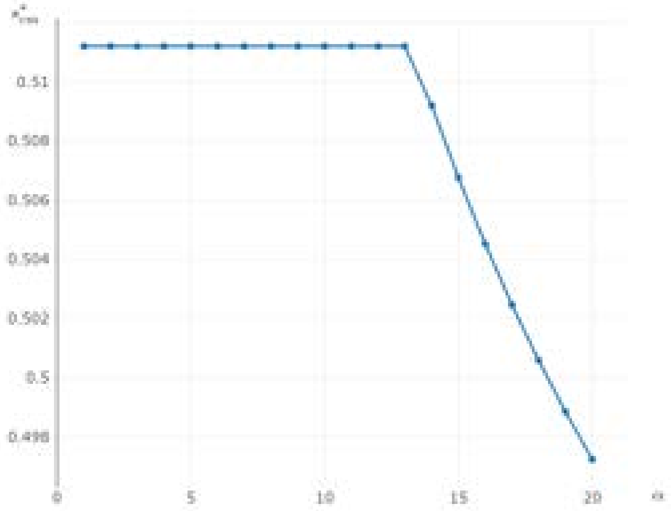

Figure 6.10: Sensitivity of s^*_{cm} to α

As shown, first s^*_{cm} does not show sensitivity to α. In addition, for $\alpha \leq 13$, patients in the system are preferred to patients who would like to enter the clinic and may dissatisfied due to being refused. However, when sensitivity to rejection is increased, the clinic should decrease the service time slightly to admit more patients as displayed in Figure 6.10.

In the following sections, clinics with quadratic and cubic arrival rate will be analyzed. Since the similar results will be obtained, explanation is omitted.

Quadratic arrival rate

Table 6.28: Different values of k,
$\theta = 14, \gamma = 4, \alpha = 1, c = 65$

k	s^*_{um}	P_{cm}	k	s^*_{cm}	P_{cm}
14	0.2276526	-0.1291908	45	0.7319004	8.809453
15	0.2439468	0.1523509	50	0.813219	10.26544
16	0.2602077	0.4347545	55	0.8945407	11.72269
17	0.2764686	0.7183685	60	0.9758626	13.18088
20	0.3252527	1.573494	61	0.992127	13.47261
25	0.4065993	3.00962	62	1.008391	13.76718
30	0.4879125	4.453589	63	1.024689	14.05691
35	0.5692299	5.902468	64	1.040953	14.34873
40	0.6505818	7.355069	65	1.044964	14.35444

Table 6.29: Different values of c,
$\theta = 14, \gamma = 4, \alpha = 1, k = 30$

c	s^*_{cm}	P_{cm}	c	s^*_{cm}	P_{cm}
30	0.4879125	4.453042	45	0.4879125	4.453276
31	0.4879125	4.453058	50	0.4879125	4.453355
32	0.4879125	4.453073	55	0.4879125	4.453433
33	0.4879125	4.453089	60	0.4879125	4.453511
35	0.4879125	4.45312	65	0.4879125	4.453589
40	0.4879125	4.453198	70	0.4879125	4.453667

Table 6.30: Different values of θ,
$\gamma = 4, \alpha = 1, k = 30, c = 65$

θ	s^*_{cm}	P_{cm}	θ	s^*_{cm}	P_{cm}
1	0.4879125	-1.720767	11	0.4879125	3.028737
2	0.4879125	-1.245817	12	0.4879125	3.503688
3	0.4879125	-0.7708662	13	0.4879125	3.978638
4	0.4879125	-0.2959158	14	0.4879125	4.453589
5	0.4879125	0.1790347	15	0.4879125	4.928539
6	0.4879125	0.6539851	16	0.4879125	5.403489
7	0.4879125	1.128936	17	0.4879125	5.87844
8	0.4879125	1.603886	18	0.4879125	6.35339
9	0.4879125	2.078836	19	0.4879125	6.828341
10	0.4879125	2.553787	20	0.4879125	7.303291

Table 6.31: Different values of γ,
$\theta = 14, \alpha = 1, k = 30, c = 65$

γ	s^*_{um}	P_{cm}	γ	s^*_{cm}	P_{cm}
1	0.4879125	6.080452	11	0.4879125	0.657574
2	0.4879125	5.538164	12	0.4879125	0.1152862
3	0.4879125	4.995876	13	0.4879125	-0.4270016
4	0.4879125	4.453589	14	0.4879125	-0.9692894
5	0.4879125	3.911301	15	0.4879125	-1.511577
6	0.4879125	3.369013	16	0.4879125	-2.053865
7	0.4879125	2.826725	17	0.4879125	-2.596153
8	0.4879125	2.284437	18	0.4879125	-3.138441
9	0.4879125	1.74215	19	0.4879125	-3.680728
10	0.4879125	1.199862	20	0.4879125	-4.223016

Table 6.32: Different values of α,
$\theta = 14, \gamma = 4, k = 25, c = 50$

α	s^*_{cm}	P_{cm}	α	s^*_{cm}	P_{cm}
1	0.4879125	4.453589	11	0.484616	4.191799
2	0.4879125	4.427022	12	0.4830769	4.17006
3	0.4879125	4.400456	13	0.4816521	4.149908
4	0.4879125	4.37389	14	0.4803146	4.131141
5	0.4879125	4.347323	15	0.4791142	4.113593
6	0.4879125	4.320757	16	0.4779466	4.097126
7	0.4879125	4.294191	17	0.4768856	4.081621
8	0.4879125	4.267625	18	0.4758838	4.066978
9	0.4879125	4.241058	19	0.4749293	4.053113
10	0.4863116	4.215371	20	0.4740276	4.039952

Cubic arrival rate

Table 6.33: Differentvalues of k,
$\theta = 14, \gamma = 4, \alpha = 1, c = 60$

k	s^*_{um}	P_{cm}	k	s^*_{cm}	P_{cm}
12	0.2186774	-0.2650416	45	0.8202156	10.39323
13	0.2369005	0.04808283	50	0.9113506	12.02566
14	0.2551569	0.3631488	51	0.9295776	12.3523
15	0.27338	0.6794828	52	0.9478046	12.67899
20	0.3645013	2.276255	53	0.9660317	13.00573
25	0.4556599	3.888366	54	0.9842587	13.33251
30	0.546788	5.508472	55	1.002486	13.65933
35	0.6379527	7.134054	56	1.020746	13.98692
40	0.729084	8.762508	57	1.038965	14.31363

Table 6.34: Different values of c,
$\theta = 14, \gamma = 4, \alpha = 1, k = 27$

c	s^*_{cm}	P_{cm}	c	s^*_{cm}	P_{cm}
27	0.49211	4.535066	45	0.49211	4.535383
28	0.49211	4.535084	50	0.49211	4.53547
29	0.49211	4.535102	55	0.49211	4.535558
30	0.49211	4.535119	60	0.49211	4.535646
35	0.49211	4.535207	65	0.49211	4.535733
40	0.49211	4.535295	70	0.49211	4.535821

Table 6.35: Different values of θ,
$\gamma = 4, \alpha = 1, k = 27, c = 60$

θ	s^*_{cm}	P_{cm}	θ	s^*_{cm}	P_{cm}
1	0.492115	-1.675163	11	0.49211	3.102382
2	0.49211	-1.197408	12	0.49211	3.580137
3	0.49211	-0.7196536	13	0.49211	4.057891
4	0.49211	-0.2418992	14	0.49211	4.535646
5	0.49211	0.2358553	15	0.49211	5.0134
6	0.49211	0.7136098	16	0.49211	5.491155
7	0.49211	1.191364	17	0.49211	5.968909
8	0.49211	1.669119	18	0.49211	6.446664
9	0.49211	2.146873	19	0.49211	6.924418
10	0.49211	2.624628	20	0.49211	7.402173

Table 6.36: Different values of γ,
$\theta = 14, \alpha = 1, k = 27, c = 60$

γ	s^*_{um}	P_{cm}	γ	s^*_{cm}	P_{cm}
1	0.49211	8.994982	11	0.49211	3.685618
2	0.49211	8.464045	12	0.49211	3.154681
3	0.49211	7.933109	13	0.49211	2.623745
4	0.49211	7.402173	14	0.49211	2.092808
5	0.49211	6.871236	15	0.49211	1.561872
6	0.49211	6.3403	16	0.49211	1.030936
7	0.49211	5.809363	17	0.49211	0.4999991
8	0.49211	5.278427	18	0.49211	-0.03093727
9	0.49211	4.74749	19	0.49211	-0.5618737
10	0.49211	4.216554	20	0.49211	-1.09281

Table 6.37: Different values of α,
$\theta = 14, \gamma = 4, k = 27, c = 60$

α	s_{cm}^*	P_{cm}	α	s_{cm}^*	P_{cm}
1	0.49211	4.535646	11	0.4871749	4.252886
2	0.49211	4.506474	12	0.4856439	4.230669
3	0.49211	4.477303	13	0.4842119	4.210016
4	0.49211	4.448131	14	0.4829142	4.190737
5	0.49211	4.41896	15	0.4816824	4.172672
6	0.49211	4.389789	16	0.4805399	4.155687
7	0.49211	4.360617	17	0.4794531	4.139667
8	0.49211	4.331446	18	0.4784297	4.124515
9	0.4906462	4.303001	19	0.477463	4.110147
10	0.4888257	4.276902	20	0.4765582	4.096491

Government's budget

In this section, it is investigated that how Government's intervention affect the performance of walk-in clinics.

As mentioned in Chapter 4, in capacitated walk-in clinic the maximum revenue is obtained at s_m.

$$R(s_m) = \overline{\lambda}\,(1 - \pi_c(s_m)) = 48.9\,(1 - \pi_c(0.23377)) = 48.8999999.$$

Assume the maximum budget assigned by the government to this plan is 1. Based on Proposition 4.4.3, there is a s_d such that

$$\pi_c(s_d) = 1 - [(1 - d)(1 - \pi_c(s_m))] = 0.0204498$$

where

$$d = \frac{B}{R(s_m)} = \frac{1}{48.9} = 0.0204499.$$

Using numerical methods, s_d is obtained:

$$s_d \approx 0.4979.$$

In Chapter 4, ancillary cost and net profit functions were obtained as below:

$$C_1(s) = p_1 R(s) = p_1[\overline{\lambda}(1 - \pi_c(s))]$$

$$N(s) = R_N(s) - C_1 - C_2 = (1 - p_1)R(s) + \frac{s}{s^*_{cm}} b_2 R(s_m) - C_2$$

Let us consider $p_1 = 0.03$. Proposition 4.4.5 represents the minimum service time to obtain more net profit. In other words, it emphasizes that to gain more net profit, the mean service time allocated to patients should be

$$s \geq s^*_{cm}(1 - p_1). \tag{6.4}$$

However, based on the limitation on budget in (4.31), there is a cap on s:

$$s \leq s_d. \tag{6.5}$$

Therefore, clinics first evaluate s_d and $s^*_{cm}(1 - p_1)$ and then start to increase service time if

$$s_d \geq s^*_{cm}(1 - p_1).$$

Since

$$0.4979 \geq 0.5111928^*(1 - 0.03) = 0.495857,$$

service time can be increased in this case. In addition, based on proposition 4.4.6, the maximum net profit is obtained at s_d:

$$N(s_d) = (1 - p_1)R(s_d) + \frac{s}{s^*_{cm}} b_2 R(s_m) - C_2$$

$$= (1 - p_1)R(s_d) + \frac{s}{s^*_{cm}}[R(s_m) - R(s_d)] - C_2$$

$$= (1 - p_1)\overline{\lambda}(1 - \pi_c(s_d)) + \frac{s_d}{s^*_{cm}}[\overline{\lambda}(\pi_c(s_d) - \pi_c(s_m))] - C_2$$

$$= (1 - 0.03)48.9(1 - \pi_c(0.4979)) + \frac{0.4979}{0.5111928}[48.9(\pi_c(0.4979) - \pi_c(0.23377))] - C_2$$

$$= 46.48108 + 0.9558373 - C_2 = 47.43692 - C_2.$$

On the other hand, minimum service time will lead to less net profit:

$$N(s_m) = (1 - p_1)\, R(s_m) - C_2$$

$$= (1 - p_1)\, \overline{\lambda}(1 - \pi_c(s_m)) - C_2$$

$$= (1 - 0.03)48.9(1 - \pi_c(0.23377)) - C_2$$

$$= 47.433 - C_2.$$

The difference between net profits is

$$N(s_d) - N(s_m) = (47.43692 - C_2) - (47.433 - C_2) = 0.00392.$$

As a result, if the governments allocate 1 budget to this plan, not only patients but also walk-in clinics will be more happy. The service time can be increased from 0.23377 to 0.4979 and net profit rises by 0.00392. However, any budget would not lead to theses satisfying results. In below flowchart, the process is shown.

In Table 6.38, the budget is considered $B = 1$. In this case, when arrival rate is 61.48 and 54.86, the budget assigned by the government can not entice the decision makers to increase the service time. In other word, when the budget is not considerable compared to the revenue gained by the patients, walk-in clinics prefer to allocate minimum service time.

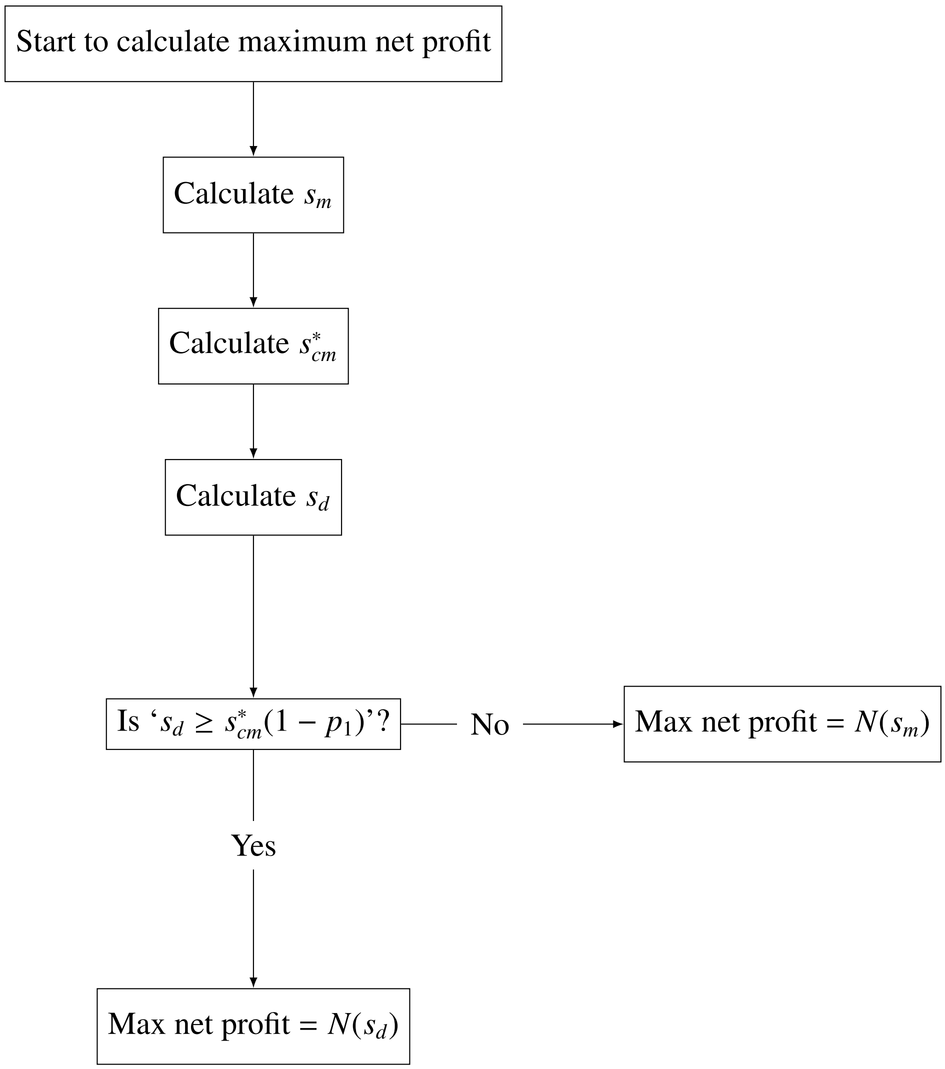
Start to calculate maximum net profit
Calculate s_m
Calculate s^*_{cm}
Calculate s_d
Is '$s_d \geq s^*_{cm}(1 - p_1)$'?
No
Max net profit = $N(s_m)$
Yes
Max net profit = $N(s_d)$

Table 6.38: Allocated service time,
$\theta = 14, \gamma = 4, \alpha = 1, k = 25, c = 50, p_1 = 0.03$

$\overline{\lambda}$	s_m	s^*_{um}	s_d	$s_d \geq s^*_{um}(1 - p_1)$	s
48.90	0.23377	0.5111928	0.4979	Yes	0.49790
61.48	0.23222	0.4065993	0.3915	No	0.23222
54.86	0.23089	0.4556599	0.4415	No	0.23089

Summary of Model CM

The main results obtained for Model CM are summarized in the followings:

- In this Scenario, clinic tries to maximize not only satisfaction of patients but also revenue.
- In chapter 3, for linear arrival rate two waiting times are obtained ($\overline{W}_{L,\lambda,l}(t)$ and $\overline{W}_{L,\lambda,l,p}(t)$). No matter which one is used as s_W, there is no significant difference between the results.
- To get positive feedback from patients, the minimum number of servers should be employed in the clinic. Considering arrival rate and other parameters (θ, γ, α and c), this number is calculated. Employing more servers will lead to increasing s^*_{um} and P_{um}.
- Impact of different parameters is shown in Table 6.39:

Table 6.39: Impact of different parameters in Model CM

Parameters	s^*_{cm}	P_{cm}
k	↗	↗
c	↗	↗
θ	↗	↗
γ	↗	↘
α	↘	↘

- Although increasing capacity results in increasing satisfaction, it is not as important as increasing the number of servers.
- Similar to Model UM, patients in the system are preferred to patients who would like to enter the clinic. However, when patients are highly sensitivity to rejection, the clinic should decrease the service time to admit more and refuse less patients.
- Comparing to sensitivity to rejection (α), when sensitivity to service time (θ) is not considerable, the number of servers and capacity should be increased. Otherwise, positive satisfaction will never be obtained even if service time decreases.

- In this scenario, clinics usually ignore the satisfaction of patients because the maximum revenue is obtained at minimum service time.

- Unlike Model UM, Government's intervention may have satisfactory results for patients and walk-in clinics. If the governments allocate enough budget to the introduced plan in this research, both patients and walk-in clinics will be more happy.

- When the budget is not considerable compared to the revenue gained by serving patients, walk-in clinics still prefer to allocate minimum service time.

6.2.3 Model UO

The only difference between Model UO and Model UM is arrival rate. Since in Model UO the clinic works in a competitive market, arrival rate depends on its performance.

Revenue

Let us assume there are 10 clinics in the area including 7 capacitated and 3 uncapacitated clinics.

$$n = n_1 + n_2 = 7 + 3 = 10$$

The information required about other competitors is provided in Table 6.40 and 6.41.

Table 6.40: Capacitated walk-in clinics

i	s_i	π_{c_i}
1	0.64	0.021
2	0.41	0.012
3	0.98	0.045
4	1.02	0.035
5	0.52	0.051
6	0.80	0.015
7	1.03	0.061

Table 6.41: Uncapacitated walk-in clinics

i	s_i
8	0.95
9	0.86

In this model, the revenue is in the form of:

$$R = \max_{s} \overline{\lambda}_{uo} = \max_{s} \left(\omega_s \overline{\lambda} + \sum_{i=1}^{n_1} \omega_s^i \pi_{c_i} \omega_i \overline{\lambda} \right),$$

$$s_m \leq s \leq s_W.$$

Based on Proposition 5.3.1, $\overline{\lambda}_{uo}$ is an increasing function of s. Therefore, maximum revenue is gained at s_W:

$$R = \max_{s} \overline{\lambda}_{uo} = \left(\omega_{s_W} + \sum_{i=1}^{n_1} \omega^{i}_{s_W} \pi_{c_i} \omega_i \right) \overline{\lambda}.$$

Maximum Revenue with different arrival rate has been calculated in Table 6.42. Also, the minimum number of servers required to allocate maximum service time is represented.

Table 6.42: Maximum revenue

$\overline{\lambda}$	s_W	ω_{s_W}	$\sum_{i=1}^{n_1} \omega^{i}_{s_W} \pi_{c_i} \omega_i$	Max R	Min k
48.9	1.052	0.1273299	0.003450248	6.395149	7
61.48	1.045	0.1265899	0.003433463	7.993836	9
54.86	1.039	0.1259547	0.003419023	7.097442	8

In this model, if the walk-in clinic decided to allocate minimum service time, then its arrival rate and consequently the revenue will reduce dramatically. The results are shown in Table 6.43 where the revenue is not comparable with revenue shown in Table 6.42.

Table 6.43: Minimum revenue

$\overline{\lambda}$	s_m	ω_{s_m}	$\sum_{i=1}^{n_1} \omega^{i}_{s_m} \pi_{c_i} \omega_i$	Min R	Min k
48.9	0.23377	0.03140479	0.000957456	1.582514	1
61.48	0.23222	0.03120306	0.0009515313	1.976864	1
54.86	0.23089	0.03102989	0.0009464432	1.754222	1

Satisfaction

In this model, patients' satisfaction depends on the average service time and the number of servers in the system. Proposition 5.3.2 shows the importance of the number of servers in satisfaction of patients when it is decided to allocate maximum service time.

6.2.4 Model CO

In this Model, we study a capaciated clinic for which the arrival rate is defined as below:

$$\overline{\lambda}_{co} = \omega_s \overline{\lambda} + \sum_{i=1}^{n_1-1} \omega_s^i \pi_{c_i} \omega_i \overline{\lambda}.$$

Let us consider 10 walk-in clinics in the region with 7 capacitated and 3 uncapacitated clinics. The information needed about other clinics are shown in Table 6.44 and Table 6.45.

Table 6.44: Capacitated clinics

i	s_i	π_{c_i}
1	0.64	0.021
2	0.41	0.012
3	0.98	0.045
4	1.02	0.035
5	0.52	0.051
6	0.80	0.015

Table 6.45: Unapacitated clinics

i	s_i
7	1.03
8	0.95
9	0.86

Revenue

In this model, revenue function is in the form of:

$$R = \max_s \left(\overline{\lambda}_{co}(1 - \pi_c^{(k)}) \right).$$

Considering different arrival rate, optimum value of service maximizing revenue is obtained depicted in Table 6.46.

Table 6.46: Maximum revenue

$\overline{\lambda}$	s_W	k	c	s_R	Max R
48.9	1.052	7	15	1.049989	5.917521
61.48	1.045	9	20	1.044964	7.647349
54.86	1.039	8	17	1.038965	6.739960

Satisfaction

In Table 6.47, optimum value of service time assuming different arrival rates is gained.

Table 6.47: Maximum satisfaction

$\overline{\lambda}$	s_W	k	c	s_{co}^*	Max P_{co}
48.9	1.052	7	15	1.051964	13.67708
61.48	1.045	9	20	1.044964	14.08003
54.86	1.039	8	17	1.038965	13.88898

Comparing results represented in Table 6.46 and in Table 6.47, no significant difference can be seen between s_R and s_{co}^*. In addition, in a competitive Market, there is no need for the government's intervention.

7. Conclusion

In this section, the main points of the thesis are reviewed and some suggestions will be given for future research.

7.0.1 Discussion and summary

In this research, non-stationary queues with time-varying arrival rate were investigated. First, ways to estimate the parameters of a non-homogeneous Poisson process were studied. Then, waiting times were estimated by using time-varying Little's Law (TVLL). Furthermore, when waiting times cannot be observed directly, Little's law can be used to estimate the average waiting time by the average number in system divided by the average arrival rate. However, applications of Little's Law (LL) with actual system data involve measurements over a finite-time interval and that simple indirect estimator tends to be biased significantly when the arrival rates are considered a time-varying function. Considering some general structural results and some simple formulas describing the time dependent performance of the IS queues with a non-homogeneous Poison arrival process, TVLL was applied to estimate waiting times.

Since it was assumed that there is appropriate time-varying staffing, the waiting time distribution was fixed even though the arrival rate was considered a time-varying function. Hence, under that condition, the TVLL provides estimation of waiting times, given estimates of the average number in system over a sub-interval and the arrival rate function. Useful variants of the TVLL estimator were obtained by fitting a linear, quadratic and cubic function to arrival data. When the arrival rate function is approximately linear, quadratic or cubic, the mean waiting time satisfies a quadratic, cubic or a polynomial of degree four equation, respectively. The new estimator based on the TVLL is a positive real root of that equation. The new methods are shown to be effective in estimating the bias in the indirect estimator.

As mentioned, fixed distribution was considered for waiting times throughout the measurement interval which was achieved by using time-varying servers as the the arrival rates change. However, in many systems such as healthcare centers which are under strain from staff shortages, the number of service providers cannot be increased as much as required. In this situation, quality of service may be sacrificed. For example, in walk-in clinics where the revenue is gained from the number of admitted patients, it would be in the best interest of clinics to reduce the service times which leads to a gap between the service a patient receives and what s/he expects. For this matter, walk-in clinics with time-varying arrival rates were studied. Based on the clinic's capacity (finite or infinite) and position of the clinic in the region (monopolistic or oligopoly), we considered four models: Model UM, Model CM, Model CO, and Model UO. Waiting times obtained in an ideal system was considered as the desired waiting time and regarding available resources, optimum value for service time was obtained in each model.

7.0.2 Future research

Future researchers can investigate the following potential opportunities:

- The initially stated overarching aim of this research was to study non-stationary queue systems in walk-in clinics and we considered polynomial functions to approximate arrival rates. However, other types of functions may also be considered for arrival rates. For instance, functions which are commonly used to model periodic phenomena such as $\sin x$ and $\cos x$.

- Another motivating extension is considering time-varying service times in walk-in clinics. The model $M_t/M_t/k$ may be an appropriate model to apply.

- In this research, we assumed exponential distributions for time-between-arrival and service time. In different walk-in clinics, different inter-arrival and service time distributions can be considered; for example, H_2 and E_4.

- Other numerical methods and techniques can be used for estimating the polynomial roots, such as Bisection, Secant, and False-Position.

- Different variables which contribute to patients' satisfaction can be considered.

- Sensitivity can be measured by statistical methods.

- Data is not from real-world cases and is generated and simulated. Real data can be applied instead of simulated and made-up data.

Bibliography

[1] Frédéric Bielen and Nathalie Demoulin. Waiting time influence on the satisfaction-loyalty relationship in services. *Managing Service Quality: An International Journal*, 2007.

[2] Vikas Kumar, Luciano Batista, and Roger Maull. The impact of operations performance on customer loyalty. *Service Science*, 3(2):158–171, 2011.

[3] William A Massey. The analysis of queues with time-varying rates for telecommunication models. *Telecommunication Systems*, 21(2):173–204, 2002.

[4] E Brockmeyer, HL Halstrøm, and A Jensen. : The life and works of ak erlang. transactions of the danish academy of technical sciences, vol. 2. the copenhagen telephone company. 1948.

[5] Michael H Rothkopf and Shmuel S Oren. A closure approximation for the nonstationary m/m/s queue. *Management Science*, 25(6):522–534, 1979.

[6] G.F. Newell. *Applications of queueing theory*. Chapman and Hall, London, 1982.

[7] William Alfred Massey. Non-stationary queues. 1982.

[8] W Randolph. Hall, queueing methods for services and manufacturing. *Englewood Cliffs: Prentice Hall*, 5:28, 1991.

[9] William A Massey and Ward Whitt. Unstable asymptotics for nonstationary queues. *Mathematics of Operations Research*, 19(2):267–291, 1994.

[10] Ira Gerhardt and Barry L Nelson. Transforming renewal processes for simulation of nonstationary arrival processes. *INFORMS Journal on Computing*, 21(4):630–640, 2009.

[11] Barry L Nelson and Ira Gerhardt. Modelling and simulating non-stationary arrival processes to facilitate analysis. *Journal of Simulation*, 5(1):3–8, 2011.

[12] Xiaowei Zhang, L Jeff Hong, and Jiheng Zhang. Scaling and modeling of call center arrivals. In *Proceedings of the Winter Simulation Conference 2014*, pages 476–485. IEEE, 2014.

[13] William A Massey, Geraldine A Parker, and Ward Whitt. Estimating the parameters of a nonhomogeneous poisson process with linear rate. *Telecommunication systems*, 5(2):361–388, 1996.

[14] William A Massey and Ward Whitt. Stationary-process approximations for the nonstationary erlang loss model. *Operations Research*, 44(6):976–983, 1996.

[15] Alan Cobham. Priority assignment in waiting line problems. *Journal of the Operations Research Society of America*, 2(1):70–76, 1954.

[16] JDC Little and SC Graves. Building intuition, vol. 115 of international series in operations research & management science, 2008.

[17] John DC Little. A proof for the queuing formula: $l = \lambda w$. *Operations research*, 9(3):383–387, 1961.

[18] Ward Whitt. A review of $l = \lambda w$. *Queueing Systems, 9:235–268,*, 1991.

[19] Dimitris Bertsimas and Georgia Mourtzinou. Transient laws of non-stationary queueing systems and their applications. *Queueing Systems*, 25(1):115–155, 1997.

[20] Brian H Fralix and Germán Riaño. A new look at transient versions of little's law, and m/g/1 preemptive last-come-first-served queues. *Journal of Applied Probability*, 47(2):459–473, 2010.

[21] Song-Hee Kim and Ward Whitt. Estimating waiting times with the time-varying little's law. *Probability in the Engineering and Informational Sciences*, 27(4):471–506, 2013.

[22] Joseph Abate and Ward Whitt. Transient behavior of the m/m/l queue: Starting at the origin. *Queueing systems*, 2(1):41–65, 1987.

[23] Joseph Abate and Ward Whitt. Transient behavior of the m/m/1 queue via laplace transforms. *Advances in Applied Probability*, 20(1):145–178, 1988.

[24] Joseph Abate and Ward Whitt. Transient behavior of regulated brownian motion, i: starting at the origin. *Advances in Applied Probability*, 19(3):560–598, 1987.

[25] Joseph Abate and Ward Whitt. Transient behavior of regulated brownian motion, ii: non-zero initial conditions. *Advances in Applied Probability*, 19(3):599–631, 1987.

[26] Peter W Glynn and Ward Whitt. Indirect estimation via $l = \lambda w$. *Operations Research*, 37(1):82–103, 1989.

[27] William S Lovejoy and Jeffrey S Desmond. Little's law flow analysis of observation unit impact and sizing. *Academic Emergency Medicine*, 18(2):183–189, 2011.

[28] Song-Hee Kim and Ward Whitt. Statistical analysis with little's law. *Operations Research*, 61(4):1030–1045, 2013.

[29] Ward Whitt and Xiaopei Zhang. Periodic little's law. *Operations Research*, 67(1):267–280, 2019.

[30] Ward Whitt and Xiaopei Zhang. A data-driven model of an emergency department. *Operations Research for Health Care*, 12:1–15, 2017.

[31] Song-Hee Kim and Ward Whitt. Are call center and hospital arrivals well modeled by nonhomogeneous poisson processes? *Manufacturing & Service Operations Management*, 16(3):464–480, 2014.

[32] Song-Hee Kim, Ponni Vel, Ward Whitt, and Won Chul Cha. Poisson and non-poisson properties in appointment-generated arrival processes: The case of an endocrinology clinic. *Operations Research Letters*, 43(3):247–253, 2015.

[33] Song-Hee Kim, Ward Whitt, and Won Chul Cha. A data-driven model of an appointment-generated arrival process at an outpatient clinic. *INFORMS Journal on Computing*, 30(1):181–199, 2018.

[34] Galit B Yom-Tov and Avishai Mandelbaum. Erlang-r: A time-varying queue with reentrant customers, in support of healthcare staffing. *Manufacturing & Service Operations Management*, 16(2):283–299, 2014.

[35] Mor Armony, Shlomo Israelit, Avishai Mandelbaum, Yariv N Marmor, Yulia Tseytlin, and Galit B Yom-Tov. On patient flow in hospitals: A data-based queueing-science perspective. *Stochastic systems*, 5(1):146–194, 2015.

[36] Sheldon H Jacobson, Shane N Hall, and James R Swisher. Discrete-event simulation of health care systems. In *Patient flow: Reducing delay in healthcare delivery*, pages 211–252. Springer, 2006.

[37] Pengyi Shi, Mabel C Chou, Jim G Dai, Ding Ding, and Joe Sim. Models and insights for hospital inpatient operations: Time-dependent ed boarding time. *Management Science*, 62(1):1–28, 2016.

[38] Joshua J Fenton, Anthony F Jerant, Klea D Bertakis, and Peter Franks. The cost of satisfaction: a national study of patient satisfaction, health care utilization, expenditures, and mortality. *Archives of internal medicine*, 172(5):405–411, 2012.

[39] Tayler M Schwartz, Miao Tai, Kavita M Babu, and Roland C Merchant. Lack of association between press ganey emergency department patient satisfaction scores and emergency department administration of analgesic medications. *Annals of emergency medicine*, 64(5):469–481, 2014.

[40] David A Gross, Stephen J Zyzanski, Elaine A Borawski, Randall D Cebul, and Kurt C Stange. Patient satisfaction with time spent with their physician. *Journal of Family Practice*, 47(2):133–138, 1998.

[41] Edwin D Boudreaux and Erin L O'Hea. Patient satisfaction in the emergency department: a review of the literature and implications for practice. *The Journal of emergency medicine*, 26(1):13–26, 2004.

[42] Edwin D Boudreaux, Roy D Ary, Cris V Mandry, and Bhrett McCabe. Determinants of patient satisfaction in a large, municipal ed: the role of demographic variables, visit characteristics, and patient perceptions. *The American journal of emergency medicine*, 18(4):394–400, 2000.

[43] Jeffrey L Jackson, Judith Chamberlin, and Kurt Kroenke. Predictors of patient satisfaction. *Social science & medicine*, 52(4):609–620, 2001.

[44] Mostafa Pazoki and Hamed Samarghandi. Regulating patient care in walk-in clinics. *Omega*, 99:102200, 2021.

[45] Goran Petrović, Nikola Petrović, and Zoran Marinković. Application of the markov theory to queuing networks. *Facta universitatis-series: Mechanical Engineering*, 6(1):45–56, 2008.

[46] Zohar Feldman, Avishai Mandelbaum, William A Massey, and Ward Whitt. Staffing of time-varying queues to achieve time-stable performance. *Management Science*, 54(2):324–338, 2008.

[47] Otis B Jennings, Avishai Mandelbaum, William A Massey, and Ward Whitt. Server staffing to meet time-varying demand. *Management Science*, 42(10):1383–1394, 1996.

[48] Stephen G Eick, William A Massey, and Ward Whitt. The physics of the mt/g/ queue. *Operations Research*, 41(4):731–742, 1993.

[49] Raymond G Ayoub. Paolo ruffini's contributions to the quintic. *Archive for history of exact sciences*, pages 253–277, 1980.

[50] A Douglas. Essentially follows clarke. *Foundations of Analysis*, page 284, 1971.

[51] Mohammadkarim Bahadori, Ehsan Teymourzadeh, Ramin Ravangard, Ali Nasiri, Mehdi Raadabadi, and Khalil Alimohammadzadeh. Factors contributing towards patient's choice of a

hospital clinic from the patients' and managers' perspective. *Electronic physician*, 8(5):2378, 2016.

[52] Crispin Jenkinson, John S Burton, Julia Cartwright, Helen Magee, Ian Hall, Chris Alcock, and Sherwood Burge. Patient attitudes to clinical trials: development of a questionnaire and results from asthma and cancer patients. *Health Expectations*, 8(3):244–252, 2005.

[53] V Merle, J-M Germain, M-P Tavolacci, C Brocard, C Chefson, C Cyvoct, S Edouard, L Guet, E Martin, and P Czernichow. Influence of infection control report cards on patients' choice of hospital: pilot survey. *Journal of hospital infection*, 71(3):263–268, 2009.

[54] Zethembiso C Hlongwa and Saajida Mahomed. Factors influencing patients' choice of clinic at inanda, kwazulu-natal. *African Journal of Primary Health Care & Family Medicine*, 13(1):2968, 2021.