UNIL | Université de Lausanne

Unicentre
CH-1015 Lausanne
http://serval.unil.ch

Year : 2012

BEHAVIOURAL QUEUEING: CELLULAR AUTOMATA AND A LABORATORY EXPERIMENT

DELGADO ALVAREZ Carlos Arturo

UNIL | Université de Lausanne

FACULTÉ DES HAUTES ÉTUDES COMMERCIALES

DÉPARTEMENT DES OPÉRATIONS

BEHAVIOURAL QUEUEING: CELLULAR AUTOMATA AND A LABORATORY EXPERIMENT

THÈSE DE DOCTORAT

présentée à la

Faculté des Hautes Etudes Commerciales
de l'Université de Lausanne

pour l'obtention du grade de
Docteur en Systèmes d'Information

par

Carlos Arturo DELGADO ALVAREZ

Directeur de thèse
Prof. Ann Van Ackere

Jury

Prof. Ghislaine Cestre, Présidente
Prof. Marco Tomassini, Expert interne
Prof. Célia Glass, Expert externe
Prof. Erik Larsen, Expert externe

LAUSANNE
2012

UNIL | Université de Lausanne
HEC Lausanne
Le Doyen
Bâtiment Internef
CH-1015 Lausanne

IMPRIMATUR

Sans se prononcer sur les opinions de l'auteur, la Faculté des Hautes Etudes Commerciales de l'Université de Lausanne autorise l'impression de la thèse de Monsieur Carlos Arturo Delgado Alvarez, licencié en Ingénierie Industrielle de l'Universidad Nacional de Colombia (Colombie) et titulaire d'un Master en Sciences en Ingénierie des Systèmes de l'Universidad Nacional de Colombia (Colombie), en vue de l'obtention du grade de docteur en Systèmes d'Information.

La thèse est intitulée :

BEHAVIOURAL QUEUEING:
CELLULAR AUTOMATA AND A LABORATORY EXPERIMENT

Lausanne, le 19 juin 2012

Le doyen

Daniel Oyon

MEMBRES DU JURY

Professeure Ann VAN ACKERE, Faculté des Hautes Etudes Commerciales, Université de Lausanne. Directeur de Thèse.

Professeur Marco TOMASSINI, Faculté des Hautes Etudes Commerciales, Université de Lausanne. Expert Interne.

Professeure Célia GLASS, Faculty of Actuarial Science and Insurance, Cass Business School, City University London. Expert Externe.

Professeur Erik LARSEN, Facoltà di scienze economiche, Institute of Management, Università della Svizzera Italiana. Expert Externe.

Professeure Ghislaine CESTRE, Faculté des Hautes Etudes Commerciales, Université de Lausanne. Présidente du Jury

University of Lausanne
Faculty of Business and Economics

Doctorate in Information Systems

I hereby certify that I have examined the doctoral thesis of

Carlos Arturo DELGADO ALVAREZ

And have found it to meet the requirements for a doctoral thesis.
All revisions that I or committee members made during the doctoral colloquium have been addressed to my entire satisfaction.

Signature: ______________________ Date: 1/6/2012

Prof. Ann VAN ACKERE
Thesis supervisor

University of Lausanne
Faculty of Business and Economics

Doctorate in Information Systems

I hereby certify that I have examined the doctoral thesis of

Carlos Arturo DELGADO ALVAREZ

And have found it to meet the requirements for a doctoral thesis. All revisions that I or committee members made during the doctoral colloquium have been addressed to my entire satisfaction.

Signature: M Tomassini Date: 8/06/2012

Prof. Marco TOMASSINI
Internal member of the doctoral committee

University of Lausanne
Faculty of Business and Economics

Doctorate in Information Systems

I hereby certify that I have examined the doctoral thesis of

Carlos Arturo DELGADO ALVAREZ

And have found it to meet the requirements for a doctoral thesis.
All revisions that I or committee members made during the doctoral
colloquium have been addressed to my entire satisfaction.

Signature: ______________ Date: 5/6/12

Prof. Célia GLASS
External member of the doctoral committee

University of Lausanne
Faculty of Business and Economics

Doctorate in Information Systems

I hereby certify that I have examined the doctoral thesis of

Carlos Arturo DELGADO ALVAREZ

And have found it to meet the requirements for a doctoral thesis. All revisions that I or committee members made during the doctoral colloquium have been addressed to my entire satisfaction.

Signature: ______________________ Date: 4/08/2012

Prof. Erik LARSEN
External member of the doctoral committee

SUMMARY

Queuing is a fact of life that we witness daily. We all have had the experience of waiting in line for some reason and we also know that it is an annoying situation. As the adage says "time is money"; this is perhaps the best way of stating what queuing problems mean for customers. Human beings are not very tolerant, but they are even less so when having to wait in line for service. Banks, roads, post offices and restaurants are just some examples where people must wait for service.

Studies of queuing phenomena have typically addressed the optimisation of performance measures (e.g. average waiting time, queue length and server utilisation rates) and the analysis of equilibrium solutions. The individual behaviour of the agents involved in queueing systems and their decision making process have received little attention. Although this work has been useful to improve the efficiency of many queueing systems, or to design new processes in social and physical systems, it has only provided us with a limited ability to explain the behaviour observed in many real queues.

In this dissertation we differ from this traditional research by analysing how the agents involved in the system make decisions instead of focusing on optimising performance measures or analysing an equilibrium solution. This dissertation builds on and extends the framework proposed by van Ackere and Larsen (2004) and van Ackere et al. (2010). We focus on studying behavioural aspects in queueing systems and incorporate this still underdeveloped framework into the operations management field. In the first chapter of this thesis we provide a general introduction to the area, as well as an overview of the results.

In Chapters 2 and 3, we use Cellular Automata (CA) to model service systems where captive interacting customers must decide each period which facility to join for service. They base this decision on their expectations of sojourn times. Each period, customers use new information (their most recent experience and that of their best performing neighbour) to form expectations of sojourn time at the different facilities. Customers update their expectations using an adaptive expectations process to combine their memory and their new information. We label “conservative” those customers who give more weight to their memory than to the new information. In contrast, when they give more weight to new information, we call them "reactive".

In Chapter 2, we consider customers with different degree of risk-aversion who take into account uncertainty. They choose which facility to join based on an estimated upper-bound of the sojourn time which they compute using their perceptions of the average sojourn time and the level of uncertainty. We assume the same exogenous service capacity for all facilities, which remains constant throughout. We first analyse the collective behaviour generated by the customers' decisions. We show that the system achieves low weighted average sojourn times when the collective behaviour results in neighbourhoods of customers loyal to a facility and the customers are approximately equally split among all facilities. The lowest weighted average sojourn time is achieved when exactly the same number of customers patronises each facility, implying that they do not wish to switch facility. In this case, the system has achieved the Nash equilibrium. We show that there is a non-monotonic relationship between the degree of risk-aversion and system performance. Customers with an intermediate degree of risk-aversion typically achieve higher sojourn times; in particular they rarely achieve the Nash equilibrium. Risk-neutral customers have the highest probability of achieving the Nash Equilibrium.

Chapter 3 considers a service system similar to the previous one but with risk-neutral customers, and relaxes the assumption of exogenous service rates. In this sense, we model a queueing system with endogenous service rates by enabling managers to adjust the service capacity of the facilities. We assume that managers do so based on their perceptions of the arrival rates and use the same principle of adaptive expectations to model these perceptions. We consider service systems in which the managers' decisions take time to be implemented. Managers are characterised by a profile which is determined by the speed at which they update their perceptions, the speed at which they take decisions, and how coherent they are when accounting for their previous decisions still to be implemented when taking their next decision. We find that the managers' decisions exhibit a strong path-dependence: owing to the initial conditions of the model, the facilities of managers with identical profiles can evolve completely differently. In some cases the system becomes "locked-in" into a monopoly or duopoly situation. The competition between managers causes the weighted average sojourn time of the system to converge to the exogenous benchmark value which they use to estimate their desired capacity. Concerning the managers' profile, we found that the more conservative a manager is regarding new information, the larger the market share his facility achieves. Additionally, the faster he takes decisions, the higher the probability that he achieves a monopoly position.

In Chapter 4 we consider a one-server queueing system with non-captive customers. We carry out an experiment aimed at analysing the way human subjects, taking on the role of the manager, take decisions in a laboratory regarding the capacity of a service facility. We adapt the model proposed by van Ackere et al (2010). This model relaxes the assumption of a captive market and allows current customers to decide whether or not to use the facility. Additionally the facility also has potential customers who currently do not patronise it, but might consider doing so in the future. We identify three groups of subjects whose decisions cause similar behavioural patterns. These groups are labelled: gradual investors, lumpy investors, and random investor. Using an autocorrelation analysis of the subjects' decisions, we illustrate that these decisions are positively correlated to the decisions taken one period early. Subsequently we formulate a heuristic to model the decision rule considered by subjects in the laboratory. We found that this decision rule fits very well for those subjects who gradually adjust capacity, but it does not capture the behaviour of the subjects of the other two groups.

In Chapter 5 we summarise the results and provide suggestions for further work. Our main contribution is the use of simulation and experimental methodologies to explain the collective behaviour generated by customers' and managers' decisions in queueing systems as well as the analysis of the individual behaviour of these agents. In this way, we differ from the typical literature related to queueing systems which focuses on optimising performance measures and the analysis of equilibrium solutions. Our work can be seen as a first step towards understanding the interaction between customer behaviour and the capacity adjustment process in queueing systems. This framework is still in its early stages and accordingly there is a large potential for further work that spans several research topics. Interesting extensions to this work include incorporating other characteristics of queueing systems which affect the customers' experience (e.g. balking, reneging and jockeying); providing customers and managers with additional information to take their decisions (e.g. service price, quality, customers' profile); analysing different decision rules and studying other characteristics which determine the profile of customers and managers.

RÉSUME

Dans cette thèse, nous étudions les aspects comportementaux d'agents qui interagissent dans des systèmes de files d'attente à l'aide de modèles de simulation et de méthodologies expérimentales. Chaque période les clients doivent choisir un prestataire de servivce. L'objectif est d’analyser l'impact des décisions des clients et des prestataires sur la formation des files d'attente.

Dans un premier cas nous considérons des clients ayant un certain degré d'aversion au risque. Sur la base de leur perception de l'attente moyenne et de la variabilité de cette attente, ils forment une estimation de la limite supérieure de l'attente chez chacun des prestataires. Chaque période, ils choisissent le prestataire pour lequel cette estimation est la plus basse. Nos résultats indiquent qu’il n’y a pas de relation monotone entre le degré d’aversion au risque et la performance globale. En effet, une population de clients ayant un degré d’aversion au risque intermédiaire encoure généralement une attente moyenne plus élevée qu'une population d'agents indifférents au risque ou très averses au risque.

Ensuite, nous incorporons les décisions des prestataires en leur permettant d'ajuster leur capacité de service sur la base de leur perception de la fréquence moyenne d’arrivées. Les résultats montrent que le comportement des clients et les décisions des prestataires présentent une forte "dépendance au sentier". En outre, nous montrons que les décisions des prestataires font converger l'attente moyenne pondérée vers l'attente de référence du marché.

Finalement, une expérience de laboratoire dans laquelle des sujets jouent le rôle de prestataire de service nous a permis de conclure que les délais d'installation et de démantèlement de capacité affectent de manière significative la performance et les décisions des sujets. En particulier, les décisions du prestataire, sont influencées par ses commandes en carnet, sa capacité de service actuellement disponible et les décisions d'ajustement de capacité qu'il a prises, mais pas encore implémentées.

ACKNOWLEDGEMENTS

First and foremost I would like to thank my mother Adriana and grandmothers “Concha” and Amparo, who have always been there to listen to me, support me and encourage me through thick and thin.

I would like to express my immense gratitude to my supervisor Ann van Ackere for her guidance and advice during these four years. Specially, I am very grateful to her for the opportunity she gave me to come to this amazing country and undertake a research career. Thanks also for her patience and willingness to help me and give me detailed and constructive feedback to my work. Furthermore, I would like to thank Professor Erik Larsen for his valuable suggestions and his willingness to help me when I needed it despite his limited availability. I also want to thank the members of the jury for reading this manuscript and for their valuable suggestions to improve it.

I am greatly indebted to Antonio Restrepo and his family for welcoming me, putting me up in their home, treating me as a family member as well as Antonio’s valuable assistance during my first weeks in Lausanne. Undoubtedly, this thesis could not have been started without their help.

I would like to thank Santiago Arango for having recommended me to Professor Ann van Ackere to pursue this PhD. I am also very grateful to him for his assistance in carrying out the experiment addressed in the fourth chapter of this thesis. I also thank Valérian Mercier and Gabriel Wichrowski for their help in conducting this experiment.

I am very grateful to Jaime Castañeda and Camila Ochoa for their valuable and helpful suggestions and criticisms of this work. Special thanks to Jaime who, no matter the time, was always willing to help me and discuss with me the different topics addressed in this thesis.

I wish to extend my gratitude to Mme Delapierre, Mme Aprile, Mme Delille, Mme Pasche, and Mme Pizzolante, members of the administrative staff at HEC-Lausanne. Thanks for their collaboration and assistance these four years with the administrative proceedings I required to fulfill my research and teaching tasks.

When I came to Switzerland to start this career, I left my family and friends in Colombia. However, I started to form a new friend family from the first day I arrived in this country. Thus, I would like to extend my gratitude to the people, who made my stay and life more pleasant in Europe. Thanks to Juan David Villegas, Bruna Jardim, Luis Gabriel Murillo, Gerardo Lecuona, Karthik Sankaranarayanan, Felipe Abaunza, Ricardo Velasquez, Germán García, Lieveke Prinsen, Fabienne Meier, Elodie Moreau, Quentin Schmieman, Fernando Cortes, Andrés Cardona, Adriana Mora, Cesar Lopez and Juan Echeverry.

Special thanks to my wife, Maira, who I married during my PhD, for her love, devotion, support and understanding during these four years. I also want to thank her for her valuable comments as I wrote this thesis.

Finally, I gratefully acknowledge the support from the Swiss National Science Foundation through Grants # 100014_116564 and # 100014_126584.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

Queueing systems have been extensively studied since Erlang (1909) tackled the telephone traffic problem for the first time. He is considered to be the founder of queueing theory and since then several related theories and concepts have been developed (e.g. Kendall, 1951; Bailey, 1954a, 1954b; Barrer, 1957a, 1957b; Naor, 1969; Yechiali, 1971; Dewan and Mendelson, 1990). Queueing theory can be applied to a large range of problems including communication and computer networks (Erlang, 1909; Agnew, 1976; Koole and Mandelbaum, 2002), service operation systems (Dewan and Mendelson, 1990), airport design (Koopman, 1972), vehicular traffic systems (Naor, 1969) and inventory systems (Graves, 1982).

This thesis addresses topics in behavioural queueing. In this first chapter we introduce the field, briefly describe the chosen methodologies and provide an overview of the results.

1.1 Queueing Systems

In general terms, a queueing system may be described as a process where customers arrive at a facility for service, and if there is no server available immediately, customers must wait to be served (Gross and Harris, 1998). The outputs of this process are served customers. Exogenous and endogenous factors can alter the performance of queueing systems. For instance, customers may decide not to use a facility due to the queue length. In this case the queue length is an endogenous factor and the customers' impatience is exogenous.

The term "customer" is used in this dissertation to refer to humans or items (e.g. vehicles, data, documents, among others), whose decision to use a service facility depends on humans or machines capable of registering and sharing information regarding the system behaviour. All customers are served in a facility whose capacity is determined by the number of available servers and the service rate. The "manager" is the service provider who makes decisions regarding the service capacity. The “queue” may be considered either as a number of customers waiting for service or as the backlog of work in a service system. Similarly the term “sojourn time” is used to denote the time from the moment the customer either submits a request to a service system (e.g. calls for appointment) or arrives at a service system to use it, until the service has been provided. In other words, this is the time from when a job or an activity is requested, until it has been carried out. Mathematically, the sojourn time is the waiting time plus the service time.

According to Gross and Harris (1998) queueing systems can be described by six major characteristics: arrival pattern of customers, service pattern of servers, queue discipline, waiting room (the authors refer this to as the system capacity), number of service channels, and number of service stages. *Arrival and service patterns* have several characteristics in common. The first is that both arrivals and services may be either individual (single customers) or simultaneous (batches of customers). Likewise, both patterns may be either deterministic or stochastic, and determined endogenously or exogenously. They are endogenously determined if they depend on the state of the system, e.g. the customers' decision of whether or not to join the facility depends on how many customers are queueing.

In the case of endogenous rates, the literature considers two possibilities: they may either depend on a steady-state (e.g. Dewan and Mendelson, 1990) or be state-dependent (e.g. Naor, 1969). The steady-state case assumes that the decisions are based on an equilibrium analysis of average performance measures of a system (van Ackere and Larsen 2009). State-dependent decisions are based on the current system status (Gross and Harris 1998). The arrival and service patterns may also be time-dependent, i.e. the arrival and service rates change with time (Gross and Harris, 1998).

Customers may exhibit impatience. Gross and Harris (1998) define three situations of impatient customers: customers deciding not to join the queue upon arrival (balking); customers, once in the queue, deciding to leave it (reneging); and customers deciding to switch queue in a system with several parallel service channels (jockeying). Customers' impatience is an exogenous factor. However their decision to stay, not join, or leave the queue may depend on both exogenous and endogenous factors.

Once customers are in the queue, they might be selected in various ways to be served. This depends on the *queue discipline*. The most familiar disciplines are first come, first served (FCFS) and last come, first served (LCFS). However, many other ways are possible, e.g. depending on the priority which the facility gives to its different customers.

A service can be made up of one or more operations or tasks, which the customers must go through to complete the service. Each of these operations or tasks is a *stage of service* and it may be performed by one or more servers; e.g. at a bank, there are several receptionists to address customers according to the required service. In the same way, one server may operate in one or more stages, e.g. in a fast food restaurant, an employee can take orders, prepare them and sometimes he also receives the cash.

The number of service channels refers to the number of parallel servers available at each stage, who are working simultaneously. A multichannel service system may have either only a single queue for each stage, as at banks, or it may have one queue for every server, as at the supermarket.

Finally the *waiting room* (referred to as *system capacity* in Gross and Harris (1998)) refers to the physical capacity of the system to accommodate a backlog of customers. When the number of customers waiting for service reaches this physical capacity, no further customers are allowed to enter until at least one customer is served or impatiently leaves. Considering the above, a system with limited queueing capacity may be viewed as a case of forced balking (Gross and Harris, 1998). However most theoretical models assume that queueing capacity is infinite.

Having presented the research area, we proceed to motivate the selection of the research problem.

1.2 Fundamental Motivation: Relationships among Customers, Queueing and Service Providers

Queueing problems concern a broad range of applications which have been widely addressed in several disciplines. These include economics, physics, mathematics, computer science and so on. Queueing is a fact of life that we witness daily. We all have had the experience of waiting in line for some reason and we know that it is an annoying situation. Banks, roads, post offices, and restaurants, are just some examples of queueing systems. There are also situations in which customers are not necessarily humans. For instance, customers could be a stack of files waiting at an office to be dealt with, jobs at a factory waiting to be performed or dispatched, and vehicles waiting at a garage to be repaired, among others. Still, note that behind each of these objects is a human being waiting (e.g. the final customer of the garage is the car owner)

Figure 1.1 Illustrative situation of queueing

As the adage says "time is money"; this is perhaps the best way of explaining what customers think about queueing. Queueing becomes even more annoying and costly for customers in those cases when they periodically require a service. In these cases, the past experience enables customers to estimate the sojourn time for the next time, before deciding whether or not to join the queue, which queue to join and/or at what time it is better to do it. This implies a dynamic queueing system with endogenous arrival rates which depend on the customers' expectations. For example, people, who annually take their car to the garage for the emissions tests, decide based on their experience what garage, to take their car to and when to do so. Similarly, a worker or a student, who daily has to choose an hour and/or a restaurant to have lunch at, has enough experience to choose the time and/or place that he considers less crowded.

Society and business have evolved. Customers have increasingly more alternatives to choose a service provider. Before, there often were few providers offering the same service or good and they were usually not close to each other. The customers therefore had to queue up at the only local facility. Nowadays there are more and more providers offering the same service or good and people are increasingly mobile. This competition enables customers to decide each time which is the best provider for their demands. Customers usually base this decision on several factors (e.g. price, quality, and waiting times), but in this work we focus exclusively on waiting times as the factor which determines customers’ decisions. When regular customers switch provider, this not only reduces revenues, but also affects the facility's reputation. A bad reputation impacts the decisions of potential customers to join the facility in the future owing to word-of-mouth effects.

In order to maximise his market share, the manager should adjust the service capacity of his facility depending on the backlog of customers. This implies that the service rates are also endogenously determined, like the arrival rates. In other cases the customer may not be able to observe the system state. We assume that the arrival and service rates are endogenously determined and that both are based on perceptions about the state of the system and in some cases on how confident agents are about the accuracy of these expectations.

We does analyse a dynamic and complex problem which depends on interactive decisions of all agents involved in the system. Customers and managers have conflicting interests: the customers' goal is to join the quickest queue and that of the managers is to satisfy the customers (i.e. keep waiting times reasonable). Finding the best configuration of a queueing system to improve its performance and to satisfy the objectives of both agents is the purpose of queueing theory, which attempts to determine how long customers should wait and how many people will join the queue (Gross and Harris, 1998).

In this section we have motivated the need to study the behavioural aspects of queueing problems. This is an area that has been little developed so far and is the focus of this thesis, as discussed below.

1.3 Collective Behaviour in Queueing Systems: the Research Problem

Most queueing problems are modelled assuming static conditions, and exogenous arrival and service rates. They are analysed in steady-state, despite the fact that they are dynamic and that agents' decisions depend on the state of the system. Over the last decades, some researchers have attempted to move away from these predominant assumptions of traditional queueing theory towards a more dynamic context in which agents' decisions are increasingly considered.

The present thesis goes in this direction. It was written as part of an SNF research project called "Queueing: A Behavioural approach" developed between 2007 and 2011. We seek to build on and extent the work of van Ackere and Larsen (2004) and van Ackere et al. (2010).

Our goal is to contribute to a new way to approach queueing problems by incorporating decision rules based on adaptive behaviour for both customers and managers. We focus on queueing problems with endogenous arrival and service patterns. We ignore others factors which can influence a customer's decision to seek service, such as quality of service, added value services and discounts. That is, we focus solely on the waiting time aspect.

We focus on understanding how customers and service providers adapt their behaviour to changing circumstances, and study the information structure used for this aim. Particularly, we wish to know how customers and managers make decisions in a queueing system and how these decisions affect and are influenced by the system behaviour. We attempt to identify and analyse if the dynamic and interactive decisions of agents (customers and manager), which are based on their experience and expectations, can lead to an adaptive and collective behaviour of the system.

We deviate from most of the literature in that we model dynamic queueing systems with arrival and service patterns which are deterministic and endogenously determined. We model the customer' and managers' decisions both from an individual and aggregated point of view by assuming state-dependent feedback. In this way, some assumptions of classical queueing theory are relaxed, such as that the system reaches a steady-state, that service rates are exogenous and that both service and arrival rates are stochastic. The customers' decision to return the next period to the same facility will depend on their past experience. Some examples of this kind of systems include a person who must choose a garage for the inspection of her car, a person who goes monthly to the bank to pay her bills, and a person who goes weekly to the supermarket, among others. In all these examples, the customer may choose the facility at which he wishes to be served and at what time to do so.

We study three different variants of queueing problems. We first consider service facility systems where customers routinely require service and autonomously choose between three facilities to be served. Customers use their most recent experience and that of their neighbours to form expectations about the sojourn time at the different facilities. The customers' decision of which facility to join is based on their expectations and on how risk-averse customers are regarding these expectations. We assume service facilities with exogenous service rate and identical service capacity, which remains constant.

In the second variant we relax the assumption of exogenous service rates by enabling managers to adjust the service capacity and solely consider risk-neutral customers.

The third variant considers a one-facility queueing system with current and potential customers. The service and the arrival rates are both endogenous; i.e. the manager can adjust the service capacity each period and customers decide whether or not to use the facility for service.

1.4 Related Work

Even though queueing problems have been extensively analysed since 1909 when A. K. Erlang studied the telephone traffic problem for the first time, most queueing models have been aimed at optimising performance measures rather than at understanding the behaviour of agents and the feedback of their decisions in a dynamic system. The early studies were confined to equilibrium theory (Kendall 1951). Subsequent work has focused on the design, running and performance of processes (van Ackere and Larsen, 2009). Most of these models are static; they assume that arrival and service rates are stochastic and exogenously determined and that the system has reached a steady-state (Rapoport et al., 2004). We define a model as static when it assumes that the system parameters (e.g. arrival and service rates) of the system remain constant during the period of analysis. On the contrary, a model is defined as dynamic when it assumes that the system characteristics may change over the period of analysis (this definition includes multi-period models).

The interest in studying the decisions involved in queueing problems is more recent. The seminal papers in this area are Naor (1969) and Yechiali (1971), who argued in a quantitative way the insight published by W. Leeman (1964 and 1965) and then discussed by T. Saaty (1965). They were the first to propose that queues could be reduced by controlling the arrival rate. They suggested the use of price for that purpose. Specifically, they proposed that the best way to reduce the queue size is by levying tolls upon arrival, i.e. customers must pay a price if they wish to use the facility’s services. Naor (1969) studied the impact on an M/M/1 queueing system (a system with a single server, a Poisson arrival process and an exponentially distributed service process) where the customers are able to decide whether or not to join the queue, based on the system congestion they observe. This is the most elementary form of feedback in a queueing system (van Ackere et al, 2006).

Although queueing theory is commonly known as a branch of operation research, queueing concerns many other disciplines. Particularly, the influence of waiting time on the relations between customers and managers of service systems has been widely addressed in marketing (Larson 1987, Law et al. 2004, and Bielen and Demoulin 2007), and management science (Naor 1969, Stidham Jr. 1985, Dewan and Mendelson 1990, Rump and Stidham Jr. 1998, and Haxholdt et al 2003). Marketing researchers study the existing relation between waiting time, customer satisfaction and customer loyalty (Law et al. 2004 and Bielen and Demoulin 2007). Their purpose is to provide managers with information regarding customers' attitudes which enable them to build strategies to improve service quality accordingly. Next we discuss the literature review from three different points of view depending on the methodologies used to study behavioural aspect in queueing problems. We first analyse the contributions made by the marketing field; next we present operations management's contributions and finally we analyse some experimental work.

1.4.1 Psychology of queues

In the 80's, marketing researchers began to develop a new field of study denominated "Psychology of queues". Larson (1987) is the seminal paper on this subject. In this paper, he raises some insights about the customers' queueing experiences. He identifies some factors which affect customers' perceptions about queues, such as the queueing environment, the availability of information regarding the anticipated delays, and the social injustice in queues. To quantify the latter element, he defines the concepts of slips and skips. A customer is *slipped* in the system when another customer who joined the queue later is first served, while this second customer has skipped over the first one. In that sense slipped people are the unhappy victims of skipping people, who often feel very satisfied.

Law et al (2004) and Bielen and Demoulin (2007) empirically study the influence of the waiting time on the satisfaction-loyalty relationship. Law et al (2004) formulate an empirical model to analyse the satisfaction level and the repurchase behaviour of customers in fast food outlets. They conclude that waiting times significantly affect the customers' satisfaction and accordingly influence their repurchase frequency. Bielen and Demoulin (2007) assess a set of eight hypotheses related to the determinants of waiting time satisfaction and the influence of this variable on the service satisfaction. Their results confirm that waiting time is a service satisfaction determinant, but additionally influence the relationship between customer satisfaction and loyalty. This paper summarises the marketing literature related to the relationship between waiting times and customer satisfaction and also covers the study of managers' strategies to match service capacity and demand.

1.4.2 Operations management models

Models focused on customers' decisions assume that they base their decisions on their last experience (e.g. Gallay, 2009 and Stidham, 1985 and 1992), on adaptive expectation (e.g. Rump and Stidham 1998, Zohar et al. 2002, van Ackere et al 2004, and Delgado et al. 2011a), or on preferences and prices for service (e.g. Naor 1969). According to Leeman, the use of prices in queueing problems decentralises the decision making of the system's agents. On the one hand, customers consider prices and their own preferences in order to decide when and where to require service. On the other hand, when the service provider applies a prices policy to reduce the queues, he should consider this policy in order to adjust the service capacity; i.e. a price system may be used as a policy to guide the hiring of personnel or the investment in the service facilities as the revenues enable financing additional capacity (Leeman, 1964). The work of Naor (1969) and Yechiali (1971) was generalised by several authors including Stidham (1985, 1992), Mendelson (1985), Dewan and Mendelson (1990), van Ackere (1995), Rump and Stidham (1998), Zohar et al. (2002), Haxholdt et al. (2003), van Ackere et al. (2006, 2010), among others. Stidham (1985), van Ackere (1995), and Haxholdt et al. (2003) provide a survey of the work on optimal pricing control and service capacity management in queueing system.

van Ackere and Larsen (2009) present an overview of a selection of key papers which have contributed to the literature on queueing through the study of decisions of agents involved in the system. They classify these papers according to three criteria related to the arrival and service patterns: 1) whether both rates are exogenous or endogenous; 2) whether both processes are stochastic or deterministic; and 3) whether feedback is state-dependent or is based on a steady-state analysis. We consider two additional criteria. The first one is whether the authors develop a model to describe queueing phenomena or conduct an experiment to analyse human behaviour in queueing phenomena. Additionally, if the papers concern models, we classify these as dynamic and static. Table 1.1 shows the classification of the literature according to these criteria. Most of these papers are discussed in detail by van Ackere and Larsen (2009). We will focus on the most relevant aspects of this literature, i.e. weaknesses and suggestion for future work, which are relevant for our research. We also discuss briefly the most recent papers related to dynamic and state-dependent models, as well as the papers which deal with experimental methods.

Most papers of table 1.1 seek to determine optimal policies for pricing and capacity decisions to control problems associated with congestion in service facility systems. The purpose of the pricing strategy is to control the expected arrival rate (Dewan and Mendelson, 1990) and the role of the optimal policy is to look for an equilibrium between the arrival rate and the admission price (Rump and Stidham, 1998). The models based on price setting and cost allocation depend on the form and the estimation of many functions (e.g. waiting time, expected price and admission price), which makes analytical solution quite complicated to find. These models are more useful for those problems where customers compare price and sojourn time when choosing a facility for service. Their decision of which facility to use therefore depends on different price-sojourn time pairs they have to choose. For instance, drivers must choose between the use of a toll road and another alternative.

Mendelson (1985) and Dewan and Mendelson (1990) study the decision problem faced by a service facility manager who must allocate costs to control the congestion level. They analyse the cases with fixed and variable service capacity. They model service facility systems in which the facility's users decide whether or not to join the facility by comparing their expected added-value to the expected marginal cost. Mendelson (1985) assumes the expected marginal cost as a linear function of the service price and an expected delay cost, while Dewan and Mendelson (1990) assume a nonlinear function. The expected delay cost can be interpreted as an opportunity cost which customers incur because of queueing delays. In other words, it is the value which users are willing to pay to avoid waiting. This value is determined using the expected average waiting time (determined in steady state) (Mendelson 1985). Both papers seek to find the optimal price and service capacity which maximise the expected net value of the total number of served customers. For the first problem, which assumes an exogenous and fixed service rate, the optimal price is equal to the expected delay cost of customers. In the second problem, which assumes variable service capacity (i.e. endogenous), the optimal price and capacity depend on the demand which materialises each period and the expected delay cost for that demand (Dewan and Mendelson 1990). Both models are static (i.e. the model is only valid if the system parameters do not change during the period of analysis). Dewan and Mendelson (1990) suggest that future research should be aimed at dynamic models with state-dependent arrival rates.

		Static models		**Dynamic models**		**Experiments**
		Stochastic	**Deterministic**	**Stochastic**	**Deterministic**	
Arrival and service rates exogenous		Edelson and Hildebrand (1975) Stidham (1985)			Agnew (1976) (Analytical)	
Arrival rate endogenous and service rate exogenous	**State dependent**	Naor (1969) Yechiali (1971) Edelson and Hildebrand (1975) Stidham (1985) van Ackere (1995) Sinha et al. (2010)		Boots and Tijms (1999)	van Ackere and Larsen (2004) Sankaranarayanan et al. (2011)	Rapoport et al. (2004) Seale et al (2005) Stein et al. (2007) Rapoport et al. (2010)
	Steady state	Mendelson (1985) Dewan and Mendelson (1990) van Ackere (1995)	Edelson (1971)	Stidham (1992) Rump and Stidham (1998) Zohar et al. (2002)		
Arrival and service rates endogenous	**Steady state**	Mendelson (1985) Dewan and Mendelson (1990)			Agnew (1976) (Descriptive)	
	State dependent	Ha (1998, 2001)			Haxholdt et al. (2003) van Ackere et al. (2006) van Ackere et al. (2010)	

Table 1.1 Literature overview (adapted from van Ackere and Larsen, 2009).

Subsequently, Stidham (1992) and Rump and Stidham (1998) extend the work of Dewan and Mendelson (1990) by developing a multi-period model (a dynamic model which evaluates the system steady-state in discrete time-periods). Both Stidham (1992) and Rump and Stidham (1998) assume that customers decide based on their estimates of the expected marginal cost whether or not to join the queue. Stidham (1992) applies static expectations (cf. Sterman 1989a) to enable customers to learn from their past experience, while Rump and Stidham (1998) apply adaptive expectations (cf. Nerlove 1958, Sterman 1989a). Static expectations assume that the current state of the system will persist, while adaptive expectations update each period the perception of this state using the most recent information. Rump and Stidham (1998) quote a wide range of papers dealing with static expectations. More recently, Sinha et al. (2010) apply an optimal pricing scheme of surplus capacity to control the joint problem of existing and potential customers who are differentiated according to a pre-specified service quality level.

Mendelson (1985), Dewan and Mendelson (1990), Stidham (1992), Rump and Stidham (1998), and Sinha et al. (2010) assume that steady-state is reached each period before the arrival rate is adjusted. Rump and Stidham (1998) suggest that it would be interesting to relax this assumption.

Haxholdt, et al. (2003), van Ackere and Larsen (2004), van Ackere, et al. (2006, 2010) and Sankaranarayanan, et al. (2011) study some behavioural aspects by analysing the feedback process involved in the customers' choice regarding what queue they should join in the next period. van Ackere et al. (2010) include the feedback process involved in the manager's decisions.

Haxholdt, et al. (2003) and van Ackere, et al. (2006 and 2010) analyse this feedback using system dynamics (SD), whereas van Ackere and Larsen (2004) and Sankaranarayanan et al. (2011) apply cellular automata (CA). the CA methodology captures the individual experience of customers and how they make their decision based on local information, while SD captures the average perceptions of customers and assumes decisions based on global information.

On the one hand, Haxholdt et al (2003) and van Ackere et al (2006 and 2010) model the customers' decisions to join a facility and the manager's response for increasing or decreasing capacity. This work is developed for a system with a single queue. In both cases, the customers' decision to join the facility is based on their perception of sojourn time. Current customers form their perceptions based on their previous experiences, while potential customers do so through word of mouth. Such expectations represent the global expectation of all the customers, who went through the system over a certain period.

On the other hand, van Ackere and Larsen (2004) and Sankaranarayanan et al (2011) model the customers' decisions to choose a queue within a system of parallel queues. These models differ from the SD models in that customers have information about their own experiences and that of their neighbours. That is, customer behaviour is modelled at an individual rather than aggregated level. van Ackere and Larsen (2004) model a road system in which they analyse how a finite number of commuters choose among alternative roads. The authors assume that commuters use their most recent experience to update expectations about travel times. Commuters compare these expectations to the most recent experience of their neighbours and choose the most attractive alternative. Sankaranarayanan et al. (2011) model a service system where customers must choose periodically among different facilities for service. They differ from van Ackere and Larsen (2004) in that they assume that customers update their expectations about sojourn times using both their own experience and that of their neighbours. Then, customers choose the facility with the lowest expected sojourn time.

1.4.3 Experiments with queueing systems

Rapoport et al. (2004) carry out a laboratory experiment aimed at studying customers' behaviour. They depict a queueing problem with endogenously determined arrival rates and state-dependent feedback as a non-cooperative n-person game. Subjects, playing the role of car owners who need to take their car to a garage for the emissions control, should decide each experimental trial whether to join the queue. If they choose to join it, they also should decide when to do so. The authors first analyse the strategies which lead to the equilibrium solutions for the game. Then, they perform an experimental study in order to identify if customers, making individual decisions, achieve a coordinated solution for this game in large groups. Finally, they characterise this solution and determine if it converges to the equilibrium. Each player was provided with information about his arrival time, his payoff for the trial, his cumulative payoff, his waiting time and the number of players who arrived at the same time for service. Players were not provided with information about the other members of the group. Subsequently Seale et al. (2005) extended this work to non-cooperative n-person games with complete information (i.e. including the information of the other group-members). Later Stein et al. (2007) and Rapoport et al. (2010) performed other experiments to study queueing systems with endogenous arrival rates and batch service. They analyse how customers decide whether to join a queue and when to do so.

1.5 Methodology

Most research on queueing problems has focused on the optimisation of performance measures and the equilibrium analysis of queueing systems. Traditionally, analytical modelling and simulation have been the approaches used to deal with queueing problems. Most simulation models are stochastic and just some more recent models are deterministic.

The analytical approach describes mathematically the operating characteristics of the system in terms of the performance measures, usually in "steady state" (Albright and Winston, 2012). This approach is useful for low-complexity problems for which an analytical solution can be found with few simplifying assumptions. Simulation is more appropriate for complex problems and enables modelling the problem in a more realistic way, as it requires fewer simplifying assumptions (Albright and Winston, 2012).

The analysis of queueing problems could be aimed either at optimising performance measures to improve the operating characteristics of the system without accounting for customer behaviour or at understanding the agents' behaviour through of the analysis of their decisions. Considering the complexity of the queueing problems described above, which is due to the interactive dynamic decisions of the agents, we will focus on studying the behavioural aspects of queueing problems by using deterministic simulation and experimental economics.

In Chapters 2 and 3 of this thesis, we use agent based simulation, specifically cellular automata (CA), to understand how customers adapt their behaviour to the system evolution as well as how customers and managers react to each other's behaviour. In the fourth chapter we apply experimental economics to analyse how human subjects, playing the role of a manager in a laboratory environment, decide when and by how much to adjust the capacity of a service facility.

1.5.1 Cellular Automata

The CA approach uses agent-based simulation (North and Macal 2007) and endows agents with enough computational ability to interact and share information with other agents of the system. This approach is useful for modelling problems at any abstraction level (Borshchev and Filippov, 2004). It has become an important tool for look into queueing problems, particularly traffic systems, from both a theoretical and a practical point of view.

A cellular automaton is a collection of cells arranged on a discrete lattice or grid which evolves at discrete times depending on the discrete states they can take (Ilachinski 2001). Cells represent agents interacting in a local neighbourhood (i.e. the discrete lattice). Every agent updates his state each discrete time step according to a decision rule which takes into account the neighbours' state.

The typical discrete lattices are either one or two-dimensional and in some cases they can be three-dimensional. In the case of one-dimensional lattices, they can have the shape of either a ring or a line and the neighbourhoods are formed by the cells that are located within a certain range (Ilachinski 2001). This range determines the number of neighbours each cell has on each side. When the lattices have more than one dimension, the neighbourhoods can take different geometrical shapes. For instance, the most familiar neighbourhoods in two-dimensional lattices are (i) von Neumann neighbourhoods where each cell has four neighbours located respectively above, below, to the left and to the right; and (ii) "Moore neighbourhoods" which have the shape of a checkerboard, i.e. each cell has exactly eight neighbours (North and Macal 2007).

John von Neumann (1948) is one of pioneers of the use of CA models. He introduced the first mathematical formulations of CA when considering a self-reproducing automata to attempt to explain the reductionist biology (Ilachinski 2001, Wolfram 1994). Nevertheless, Stephen Wolfram is recognised as the main contributor to the development of this field throughout the eighties and nineties (Wolfram (1994) contains a collection of their most outstanding papers on CA and complexity). CA have been used to represent and understand the behaviour in several complex systems, including physical fluids, neural networks, molecular dynamical systems, natural ecologies, military command and control networks, and the economy, among many others.

We select CA because it is a very simple method to study behavioural pattern formation in complex systems. Queueing systems are complex in the sense that they are dynamical systems where the interactions between customers as well as between customers and managers are typically non-linear. We use a CA model to depict how customers interact in a local neighbourhood to decide each period which facility to patronise for service based on their own experience and that of their neighbours. The neighbourhood represents a social network of customers with the same aim, e.g. colleagues, friends, people living next-door etc.

We apply a one-dimensional CA to depict a queueing situation in a captive market where customers routinely require service and must choose among a set of facilities to be served. The discrete lattice has the shape of a ring. Customers are the cells and the facilities represent the states each cell can take. Each cell has exactly K neighbours on each side. For instance, when $K = 1$ the last cell has the first one and the last but one as neighbours. Each period, customers must choose a facility (state) following a decision rule based on their expectations of the sojourn time at the facilities. Customers form their expectations using their most recent information. When a customer patronises a facility he uses this experience to update his expected sojourn time at this facility. Moreover, customers share their experience with their neighbours and use this information to update the expected sojourn time at the facility of their best performing neighbour. Customers update their expectations each period using an adaptive expectation process (Nerlove, 1958) whereby they weight their memory and their most recent information. The adaptive expectations concept is also known as exponential smoothing (Theil and Wage, 1964).

van Ackere and Larsen (2004) assume that customers update their expectation using solely their most recent experience and that they then compare this expectation to the latest experience of their neighbours in order to decide which road to select the next period. We deviate from van Ackere and Larsen (2004) in that we model customers who update their expectations not only using their own experience, but also that of their neighbours. Customers update the expectation of their last patronised facility using their own experience and use the most recent experience of their quickest neighbour (i.e. the best performing) to update their expectation of the facility used by this neighbour. Subsequently they compare their expectations in order to decide which queue to join. This is the basic structure of the CA model we use in the next two chapters of this thesis.

In Chapter 2 we introduce the idea that customers are uncertain about their expectations. The extent to which they account for this uncertainty when taking decisions depends on how risk-averse they are. Customers estimate their uncertainty through the volatility of the estimate of their expectations. Given the use of adaptive expectations, this volatility is usually estimated by means of the smoothing variance (Taylor 2006). The variance is unobservable, but this can be estimated by applying exponential smoothing to the squared residuals of the expectations (Taylor 2004). Customers use this estimate of uncertainty to compute an upper-bound of the sojourn time on which they base their decision of which facility to join each period. We assume exogenous service rates: facilities having identical service capacity which remains constant over time. Hence, the average sojourn time of the system only depends on the customers' decisions.

In Chapter 3 we use the same basic CA model, but we assume risk-neutral customers and relax the assumption of exogenous service rates. We enable managers to adjust the service capacity of the facilities (i.e. endogenous service rates). We assume that managers base their capacity adjustment decisions on their perception of the arrival rate. We use the same principle of adaptive expectations and endow managers with memory to update their perceptions about the future arrival rates. We model service systems where the managers' decisions require time to be implemented, e.g. hiring employees requires training, laying off staff may imply a notice period, acquiring IT systems takes time, among others. Managers are characterised by the speed at which they update their perceptions, the speed at which they take decisions, and how coherent they are when accounting for their previous decisions still to be implemented when taking their next decision.

1.5.2 Laboratory Experiments

Experimental economics is a methodology that allows collecting data from human subjects to study their behaviour in a controlled economic environment (Friedman and Sunder, 1994). Laboratory experiments were introduced in Economics by Edward H. Chamberlin (1948). Motivated by the theories of industrial organisation and market performance, he looked into the behavioural characteristics of competitive markets and tested the hypothesis that monopolistic competition theories are more useful to explain the observed behaviour than the theories of demand and supply (Plott 1982).

Models are useful to predict behaviours and explain how these behaviours are influenced by changes in independent variables, while laboratory experiments are useful to explain how these behaviours emerge. Experimental data analysis provides us with information to explain the effects some variables have on micro-economic phenomena which existing theories cannot explain (Friedman and Sunder, 1994).

The fundamental objective of experimental economics is to recreate a controlled micro-economic environment in the laboratory which guarantees the measurement of the relevant variables (Smith 1982). In this sense, the relevance of the collected data rests on the proposition that laboratory environments represent "real" markets in which economic principles can be applied when real people pursue real profits within a virtual context of real rules (Plott 1982).

The literature presents three basic principles which must be considered for the design of economic experiments: 1) realism, 2) control and repetition and 3) induced-value theory (Friedman and Sunder, 1994). These principles are a guidance for the economic experimentalists. *Realism* refers to the difference between models and experiments. A model may simulate the reality, but it performs under the conditions, rules and laws assumed by the modeller and not under conditions, rules and laws under which individuals and institutions interact in reality. Experimental economics evokes a parallel environment between reality and laboratory experiments (Grossklags 2007). *Controlling and repeating* are required to ensure that the collected observations are the result of the manipulation of the experimental environment by the experimenter (Grossklags 2007). However, in any experiment there are factors which cannot be controlled. The goal of the repetition is precisely to reduce the variability caused by these factors. *Induced-value theory* proposes the use of a reward medium to convince subjects to participate in the experiment and induce pre-specified characteristics (Friedman and Sunder 1994). This principle distinguishes experimental economics from other experimental disciplines.

In Chapter 4, we perform an experiment in which we recreate the capacity management of a service facility through a controlled laboratory environment. We use the protocols of experimental economics (Friedman and Sunder 1994) to design, carry out, and control this experiment.

Consider a situation where customers repeatedly choose a service provider. The system is composed of a service facility, a manager, a queue and customers. The experiment is presented as a garage with two populations of customers seeking service: regular customers and potential customers. Regular customers are those who currently patronise the facility and periodically decide whether or not to use its service. The frequency at which they do so is given and known by the manager. They may leave the system if their expected sojourn time (which is based on (an average of) the last few times they have used the facility's services) is greater than the time they consider acceptable. Potential customers currently do not use the facility, but might consider doing so in the future depending on how attractive the sojourn time is, i.e. they might become regular customers.

The system operates as follows: regular customers request a service which will be provide immediately if there are free servers at that moment; else, they join the queue. Customers in the queue are served according to the first-come, first-served (FCFS) rule. The service capacity is adjusted by the manager, who knows the current system state, in order to achieve an acceptable sojourn time. However, the manager's decisions are not instantaneously implemented. Indeed, capacity adjustment involves an explicit time-lag to delivery capacity orders or to dismantle capacity. There are other delays involved in the system, which are unknown for the manager. These delays are the time which customers take to update their perceptions of sojourn time.

We adapt the system dynamics (SD) model proposed by van Ackere et al (2010) to represent the above queueing situation. We develop a computational graphical user interface adapted to this model to perform our experiment. The participants of this experiment are Bachelor and Master Students from the Faculty of Business and Economics at the University of Lausanne. Each participant plays the role of the manager of a garage and his task is to adjust the service capacity. The customers of the garage are virtual agents whose behaviour is generated by the underlying model.

Our objective is to study the impact of delays on the managers' capacity decisions. We wish to test in particular whether subjects, taking on the role of managers of a service facility, account for their previous decisions when taking the next capacity decision. Additionally, we attempt to determine the decision rule used by the subjects in the laboratory to manage the capacity adjustment process.

1.6 Overview of Results Contributions

As mentioned before, this thesis is associated to the SNF research project "Queueing: A Behavioural approach". This project has already yielded eight publications, listed in Appendix A. Four of these publications are directly related to this thesis and are included in Appendices B to E. This thesis is divided into three parts: first we analyse queueing problems from the standpoint of risk-averse customers, then we incorporate the managers' decisions in order to analyse the behaviour resulting from the interactions between customers and managers, and finally we perform a laboratory experiment to collect information about how human subjects taking the role of managers adjust the service capacity.

This thesis thus tackles three different queueing situations. The first situation considers service facility systems where customers routinely require service and autonomously choose between *m* facilities to be served. Each facility has its own queue. We assume service facilities with constant exogenous service rate and identical service capacity. This issue is addressed in Delgado et al. (2011a) (see Appendix B), Delgado et al. (2011b) (see Appendix C), Delgado et al. (2011d) (see Appendix E) and Chapter 2 of this thesis.

In Delgado et al. (2011a), the authors analyse the different collective behavioural patterns which emerge from the decisions of interacting customers who routinely choose a facility for service. These patterns are: chaotic behaviour, almost-stable behaviour and the Nash Equilibrium. In the first pattern, customers never find a facility which meets their needs and therefore continue to switch facility irregularly. In the second pattern, groups of neighbours are loyal to a facility and some customer on the borders of these neighbourhoods alternate between two facilities. A special case of this pattern occurs when the customers' expectations regarding one of the facilities become very large and agents decide to never again patronise this facility. The last pattern occurs when customers are equally split among all facilities and yields the lowest possible average sojourn time. Customers are satisfied with their experience and do not wish to change facility.

In Delgado et al. (2011b), Delgado et al. (2011d) and Chapter 2 of this thesis, we consider customers whose degree of risk-aversion determines the extent to which they account for uncertainty when taking the decision of which facility to join (as mentioned in Section 1.5.1). Delgado et al. (2011b) focus on the behaviour of risk-neutral and moderately risk-averse customers. They conclude that systems with customers having an intermediate degree of risk-aversion exhibit longer transient periods and converge more slowly to an almost-stable behaviour. Delgado et al. (2011d) show that customers with a high level of risk-aversion achieve low sojourn times when they are cautious towards the new information used to update their expected sojourn time, whatever their attitude regarding the information used to update the variance. Customers with an intermediate level of risk-aversion experience low sojourn times when they are reluctant to update their expectations of both sojourn time and variance. Finally, risk-neutral customers and those with low risk-aversion achieve their best performance when they are most conservative regarding the updating of their expected sojourn times.

Chapter 2 of this thesis integrates Delgado et al. (2011b) and Delgado et al. (2011d) and analyses more thoroughly the behaviour of the risk-averse customers. We show that there is a non-monotonic relationship between the degree of risk-aversion and system performance. For instance, customers with an intermediate degree of risk-aversion typically achieve higher sojourn time than the others; moreover, it is very unusual for this type of customers to reach the Nash equilibrium. Risk-neutral customers have the highest probability of achieving the Nash Equilibrium. Concerning the transient period, we extend the conclusion of Delgado et al. (2011b) to risk-averse customers in general, i.e. the more risk-averse the customers, the longer the transient period exhibited by the system.

In Chapter 3, we consider a queueing system with similar characteristics to that addressed in Chapter 2, but now we incorporate endogenous service rates and focus solely on risk-neutral customers. We assume that both customers and managers base their decisions on their perception of the state of the system. While customers have information about the most recent sojourn time they experience at the facilities, managers know the number of arriving customers at their facilities. So both customers and managers use this information to update their perceptions. We find that the managers' decisions exhibit a strong path-dependence: managers can have the same profile, but owing to the initial conditions of the model the facilities evolve completely differently. In some cases the system becomes "locked-in" into a monopoly or duopoly situation. The weighted average sojourn time of the system converges to the exogenous benchmark value which managers use to estimate their desired capacity. We also find that the more conservative a manager is regarding new information, the larger the market share his facility achieves. Additionally, the faster he takes decisions, the higher the probability that he achieves a monopoly position.

Delgado et al. (2011c) (see Appendix D) and Chapter 4 address the laboratory experiment described in Section 1.5.2. We consider a different queueing situation to that tackled in Chapters 2 and 3. Delgado et al. (2011c) analyse the typical behaviour of the system occurring when one of the equilibrium conditions is modified. They propose two alternative strategies to manage the capacity adjustment of the service facility for which they determine the optimal parameters and analyse the resulting system behaviour. In Chapter 4, we identify three groups of subjects whose decisions result in similar behavioural patterns. These groups are labelled gradual investors, lumpy investors, and random investor. The autocorrelation analysis of the subjects' decisions indicates that these decisions are positively correlated to the decisions taken one period early. Subsequently we formulate a heuristic to model the decision rule considered by subjects in the laboratory. We find that this decision rule fits very well for those subjects who gradually adjust capacity, but it does not capture the behaviour of the subjects of the other two groups. The experiment was performed for different treatments whereby we varied the length of the delays involved in the system. The results indicate that the longer the delivery and dismantling delays, the lower the cumulative profits typically achieved by subjects in the lab. In contrast, the different delays involved in the updating process of the customers' perception do not significantly affect these profits

The main contribution of this thesis is the use of simulation and experimental methodologies to explain the collective behaviour generated by customers' and managers' decisions in queueing systems as well as the analysis of the individual behaviour of these agents. In this way, we differ from the classical literature related to queueing systems which focuses on optimising performance measures and the analysis of equilibrium solutions. Our work is a building block for further theoretical work on the capacity adjustment of service facilities implying queueing phenomena. This framework is still in its early stages and accordingly there is a large potential for further work that spans several research topics. Interesting extensions to this work include incorporating other characteristics of queueing systems, which affect the customers' experience (e.g. balking, reneging and jockeying); providing customers and managers with additional information to take their decisions (e.g. price, quality, customers' profile, among others); analysing different decision rules and studying other characteristics which determine the profile of customers and managers.

1.7 Overview of the Thesis

The next three chapters are structured as research papers consisting of a brief introduction of the topic, a description of the methodology, a discussion of the results, conclusions, and suggestions for further work.

Chapter 2 addresses a queueing system with risk-averse customers and exogenous service rates. After a brief introduction, we describe in section 2.2 the elements which make up this queueing system. Section 2.3 introduces the CA model which depicts how customers interact and share information in this system when choosing a facility for service. The information on which customers base this decision and the way they update such information are explained in Section 2.4. There, we describe the adaptive expectation process used to update the customers' perceptions of sojourn time as well as the way customers estimate the uncertainty involved in these perceptions. In Section 2.5 we characterise the customers' profile according to how risk-averse they are and the weight they give to new information when updating their perceptions and estimating uncertainty. Section 2.6 discusses the collective behaviour observed when customers are risk-neutral and risk-averse. Section 2.7 details the sensitivity analysis of the weighted average sojourn time of the system with respect to the parameters which determine the customers' profile. The last section addresses the conclusions and further work.

In Chapter 3 we relax the assumption of exogenous service rate and incorporate the service provider's decisions into the behavioural analysis of the agents involved in queueing systems. Section 3.1 gives a brief introduction of this topic. Section 3.2 describes the adaptations we have made to the CA model of Chapter 2 in order to consider endogenous service rates. In the same section we define the managers' and the customers' decision rules. Section 3.3 characterises the managers' and customers' profiles as a function of the model parameters. The next section discusses some typical behavioural patterns of customers and managers, and provides an aggregated view of the system when these patterns occur. Then we perform some experiments in which we analyse the influence which the different parameters have on the performance of the facilities. Finally, we present conclusions and suggestions for further work.

In Chapter 4 we address the laboratory experiment performed to analyse the way human subjects make decisions regarding the service capacity of a facility with current and potential customers. Section 4.1 briefly describes the SD model used in this chapter, focusing on the adaptations made to the original model developed by van Ackere et al. (2010). Section 4.3 describes the experimental economics protocol (Friedman and Sunder 1994) used to conduct this experiment. In Section 4.4 we discuss the results. We first provide a descriptive analysis of the way subjects take their decisions. Next, we test the experimental hypothesis and analyse differences between treatments. Finally we present conclusions and provide some insights for further work.

In Chapter 5 we summarise the results of this thesis and provide more general conclusions of this thesis. We also discuss the contribution of the thesis to study the behavioural patterns in queueing systems and service systems in general. Additionally, we present the limitations of the present work. As mentioned before, the appendix contains the papers we have published in the course of this research and which are related to this thesis.

2 QUEUEING SYSTEMS WITH RISK-AVERSE ADAPTIVE CUSTOMERS

ABSTRACT

Queueing problems cover a wide gamut of applications that have extensively been addressed in various disciplines. However, most research on this subject has been mainly aimed at the optimisation of performance measures and the equilibrium analysis of a queueing system. The decision making process of the customers of the facility and the impact of their individual choice on queue formation have rarely been studied. In this chapter, we tackle this process in dynamic queueing systems with deterministic endogenous arrivals. A self-organising queueing system with local interaction and locally rational customers is assumed to portray the way customers, who routinely require a service, decide which facility to use. We deviate from most of the literature in that we model the customers' decision process by applying adaptive expectations and incorporating the uncertainty involved in these expectations. Customers update their expectations based on their own experience and that of their neighbours. We use simulation analysis to compare the collective behaviour of risk-neutral and risk-averse customers. Risk-neutral customers ignore uncertainty and base their decision only on their expected sojourn time, while risk-averse customers account for uncertainty and use it to estimate an upper bound of the sojourn times. A one-dimensional cellular automata model is used to explain how customers interact in a multichannel service facility and study their collective behaviour. Our results indicate that the more risk-averse the customers the longer the transient period the system exhibits. Additionally, after this transient period, the system converges more slowly to an almost-stable average sojourn time. Systems where customers are either close to risk-neutral or strongly risk-averse perform better, in terms of average sojourn time, than those whose customers have an intermediate level of risk-aversion.

KEYWORDS: Queueing problems, cellular automata (CA), adaptive expectations, uncertainty.

2.1 Introduction

Queueing problems have been extensively studied since Erlang (1909) published his work on telephone traffic problem in 1909. These problems span a wide range of applications, which concern several disciplines such as operation research, economics, management and computer science, among others. Queueing system applications include network systems, service operations, inventory systems and traffic systems, to name a few.

A queueing system can be described as a process where customers arrive at a facility for service and they must wait when such a service is not immediate (Gross and Harris 1998). Customers can be humans or objects (e.g. vehicles, data, documents, among others) requiring service. In many cases, customers require a service routinely; they must thus decide which facility to patronise each period. Some examples of this kind of systems include a person who goes weekly to the supermarket, a person who must choose a garage for the inspection of her car, and a person who goes monthly to the bank to pay her bills. These decisions often depend on customers' previous experience at the facility.

Since Erlang (1909), queueing problems have been mainly tackled from an aggregated point of view in which researchers are focused on optimising the performance measures of the system (e.g. average sojourn time). Most of the early works concerning queueing problems were confined to the equilibrium theory (Kendall 1951). This work has been useful to improve the performance of many systems, or in other instances to design new processes in social and physical systems. However, in many cases this view has also limited our ability to explain the behaviour observed in many real queues.

Koole and Mandelbaum (2002) emphasise the need to include human factors in the context of queueing models, particularly in cases such as call centres. Later, Sankaranarayanan et al. (2011) generalise this thought and suggest delving into micro-dynamics of queueing systems and analysing how the queues are formed. Nevertheless, in the 80's, marketing researchers began to develop a new field of study denominated "Psychology of queues". Larson (1987) is a seminal paper in this subject. In this field, researchers attempt to understand how customers decide "to queue or not to queue". More recently marketing researchers have shifted their focus toward the study of the influence of the waiting times on customer satisfaction, customer loyalty and service quality (Law, et al. 2004; Bielen and Demoulin 2007).

Other disciplines, such as management science, have focused on building models to study customers' decisions (e.g. Dewan and Mendelson 1990; Rump and Stidham Jr. 1998; van Ackere and Larsen 2004; Sankaranarayanan, et al. 2011). In some of these models, customers' decisions are based on customers' expectation of sojourn time (Zohar, et al. 2002; Haxholdt et al. 2003). Other authors model these decisions by incorporating a price as mechanism to control the arrival rates (e.g. Naor 1969; Stidham Jr. 1985; Stidham Jr. 1992; Dewan and Mendelson 1990). This stream of literature started when Naor (1969) and Yechiali (1971) published their seminal papers in which they argue in a quantitative way the insight published by W.A. Leeman (1964) and then discussed by T. L. Saaty (1965) and W.A. Leeman (1965). They are the first to propose that queues could be reduced by controlling the arrival rate. Specifically, they propose that the best way to reduce the queue size is by levying tolls upon arrival, i.e. customers must pay a price if they wish to use the facility's services.

Most research, which has continued with this stream (i.e. incorporating customers' decisions into the queueing model), considers queueing systems with stochastic arrival and service patterns. The resulting models have been studied theoretically in operations research and management science by Edelson and Hilderbrand (1975), Stidham Jr. (1985), Dewan and Mendelson (1990), van Ackere (1995), Rump and Stidham Jr. (1998), Zohar et al. (2002), among others. Hassin and Haviv (2003) widely discuss the literature on equilibrium behaviour of customers and servers in stochastic queueing systems. Comparatively, deterministic models have not been much discussed. Some models were proposed by Edelson (1971), Agnew (1976), Haxholdt et al. (2003), van Ackere and Larsen (2004), and van Ackere et al. (2006).

Haxholdt et al. (2003) and van Ackere et al. (2006), using system dynamics, include feedback into their model to look at the return rates of customers to the service facility. van Ackere and Larsen (2004) apply a one-dimensional CA model in order to study the formation of commuters' expectations about congestion in a three-road system. Commuters update their perceptions using an exponentially weighted average. Then, by comparing these perceptions to the most recent experience of their neighbours, commuters pick the best alternative regarding the expected travel time.

Later, Sankaranarayanan et al. (2011) and Delgado et al. (2011a) (see Appendix B) use a similar approach to model a multichannel service facility in which customers routinely choose a service facility for service. Differing from van Ackere and Larsen (2004), Sankaranarayanan et al. (2011) and Delgado et al. (2011a) assume that customers update their expectations using both their own information and that of their neighbours. In this case, a different weight is given to the information provided by the neighbours. Delgado et al. (2011a) explain how local interactions among customers influence the formation of queues and the different collective behaviours patterns, which the system exhibits Sankaranarayanan et al. (2011) delve into the state of the art of deterministic models applied to understanding the way customers behave in queueing systems and control the arrival rates.

In this chapter, we propose an extension of the model proposed by Sankaranarayanan et al. (2011) and Delgado et al. (2011a). Our work differs from the above papers in the following aspects: (i) we use the concept of volatility of forecast errors (Taylor 2004 and 2006) to incorporate uncertainty into the formation process of expectations proposed by Sankaranarayanan et al. (2011); (ii) we differentiate between risk-neutral and risk-averse customers; (iii) we analyse how different levels of risk-aversion may affect the collective behaviour of customers; and (iv) we assume that customers give the same weight to their own information and that provided by their neighbours, but they apply different weights to estimate their expectations regarding sojourn times and their uncertainty.

Our simulation results indicate that the more risk-averse the customers are, the longer the transient period exhibited by the system is. After this transient period, if customers are risk-averse, the system converges more slowly to an almost-stable average sojourn time. Systems where customers are either risk-neutral or strongly risk-averse perform better and are less likely to ignore a facility (i.e. never use it again) than those whose customers have an intermediate level of risk-aversion.

The remainder of the chapter is structured as follows. The next section describes the system we deal with in this chapter. Section 2.3 introduces the CA model and provides a technical description. Then, we present the concept of adaptive expectations and explain how we use it to estimate the expected sojourn times and the customers' uncertainty regarding these expectations. Customers use the resultant estimates in order to form their own measure of the system performance, which enables them to decide which facility to join each period. The fifth section of this chapter characterises the customer types according to their risk-aversion attitude and the weights they give to their expectations. Later, we discuss and explain the simulation results for a typical case. In Section 2.7, we present a sensitivity analysis with respect to the risk-aversion parameter and the expectation coefficients. The last section addresses the conclusions and future work.

2.2 A Service System

Consider a fixed population of N interacting and homogenous customers requiring service. All customers need service simultaneously and it is compulsory for them to choose one of m facilities. This system portrays a captive market in which customers repeatedly need either a service or a good and they have several options to obtain it. Examples include food stores, post offices, garages and banks, among others. The assumption that customers need service simultaneously is a stylised representation of a rush-hour. Thus, a queue will form at each facility when the rate of arriving customers exceeds the facility's service capacity. Queues are unobservable at the time of decision making and decisions are irreversible: customers do not have any accurate information regarding the queue size or sojourn time when making their choice and cannot switch facility after making their choice. Nonetheless, they can form perceptions of the sojourn times at their previously chosen facilities based on their experience. Additionally, according to the aforementioned attribute of interacting customers, we assume they interact with their neighbours and share information about their previous experience. Hence, customers update their perceptions of sojourn time for the facilities, which they and their best performing neighbours patronised the previous period. The customers' decision of which facility to join therefore depends on their perceptions.

2.3 The CA Model

In this chapter the system described above is modelled as a queueing system with endogenous arrival rates and exogenous service rates. We assume identical service rates for all facilities, while arrivals at each facility depend on the customers' previous experiences and their expectations. Reneging and balking are not allowed.

We adopt a CA approach (North and Macal 2007; Wolfram 1994) to model the interaction between customers, capture their expectations and analyse their collective behaviour. In the context of the CA methodology (North and Macal 2007), agents portray customers. They interact in a one-dimensional neighbourhood assumed to have the shape of a ring and composed of cells. Each cell is an agent who has exactly *K* neighbours on each side. For instance, when *K = 1,* agent *i* has agents *i-1* and *i+1* as neighbours. The last and the first agents located in the ring are special cases. The first agent has the second agent and the last agent as neighbours, while the neighbours of the last agent are the penultimate agent and the first one. The neighbourhood can depict for example a social network encompassing colleagues, friends, people living next-door etc. The facilities represent the states that the cells (agents) can take each period. Each period, agents must choose a facility (state) following a decision rule based on their most recent experience. We assume smart agents who have the ability to remember information and update their memory using new information. The updating process of agents' memory is based on the theory of adaptive expectations (Nerlove 1958). This theory is also known as exponential smoothing and assumes that agents use the most recent information, which they have, to adapt their expectations. This information stems from their last experience and that of their best performing neighbour, i.e. the neighbour who has experienced the minimum sojourn time in the previous period. Moreover, these expectations involve a certain degree of uncertainty which is captured by the variance of the estimates. This variance is computed by using the squared errors of the forecasts. As variance is unobservable, exponential smoothing is applied to estimate the squared residuals (Taylor 2006). This method is known as volatility of forecasting errors (e.g. Taylor 2004).

Technically the CA model works in the following way. Let $\mathcal{A}$ be a set of *n* agents (cells) $\{A_1, A_2,\ldots, A_i,\ldots, A_n\}$ and $\mathcal{Q}$ the set of *m* facilities (states) $\{Q_1, Q_2,\ldots, Q_j,\ldots, Q_m\}$ which agents can choose (take). Each agent A_i must join exactly one facility Q_j each period *t*. All *m* facilities have the same service rate μ, but different arrival rates (λ_{jt}). We define the arrival rate as a function of the state of the agents each period, $s_i(t)$. The agents' decisions will determine their state for each period. Let $\mathcal{S}$ denotes the set of states $s_i(t)$ of *n* agents in period *t.* This state $s_i(t)$ is one of the *m* possible facilities, that is, $\mathcal{S} \subset \{Q_1, Q_2,\ldots, Q_j,\ldots, Q_m\}$. With this in mind, the arrival rate (λ_{jt}) for queue *j* at time *t* can be written as a function of S, *Q,* and *t,* given by the following equations:

$$x_{ij}(t) = f(s_i, Q_j, t) = \begin{cases} 1 & \mathit{if}\, s_i(t){=}Q_j \\ 0 & \mathit{otherwise} \end{cases} \qquad (2.1)$$

$$\lambda_{jt} = \sum_{i=1}^{n} x_{ij}(t) \qquad (2.2)$$

We assume that the state $s_i(t)$ evolves over time depending on the agent's expected sojourn times, his uncertainty and his risk-aversion. Let us denote the expected sojourn time of agent A_i for the facility Q_j in period *t* by M_{ijt}, the corresponding uncertainty by σ_{ijt}, and the risk-aversion factor by R. M_{ijt} and σ_{ijt} evolve over time, while R is assumed to be a constant. Additionally, this parameter is the same for all agents, as implied by the aforementioned assumption of homogenous customers. Then, the state $s_i(t)$ for agent A_i will be determined through the following function:

$$s_i(t+1) = F(M_{ijt+1}, \sigma_{ijt}, R) \tag{2.3}$$

In order to define this function, we incorporate these variables into a measure that enables agents to decide their state each period. We call this measure *the expected upper bound of the sojourn time* and denote it by U_{ijt}. By upper bound, we mean the maximum sojourn time, which agents estimate they may experience at the facilities. This upper bound is estimated by using the agents' expected sojourn time (M_{ijt}) and their uncertainty and risk-aversion level. Finally, assuming that agents are rational, in the sense that they always choose whatever is most convenient for them, they will decide to patronise the facility with the lowest upper bound. Before delving into the function that determines the expected upper bound of sojourn time (U_{ijt}), we describe how agents form their expectations and estimate their uncertainty by applying an adaptive expectation model.

2.4 The Adaptive Expectations Model

Agents update their expected sojourn time M_{ijt} and their uncertainty σ_{ijt} by applying adaptive expectations (Nerlove 1958), also called adaptive or exponential forecasting (Theil and Wage 1964; Gardner Jr. 2006). This is a mathematical-statistical method of forecasting commonly applied to financial market and economic data, but it can be used with any discrete set of repeated measurements (Gardner Jr. 2006). This technique is based on the weighted average of two sources of evidence: the latest evidence (the most recent observation), and the value computed one period before (Theil and Wage 1964).

2.4.1 Estimating the Expected Sojourn Times

Considering the assumption of captive customers, the latest evidence they have to estimate their expected sojourn time is given by their most recent experience in the system. This experience is denoted by W_{ijt}. Thus, agent A_i, who uses facility Q_j in period *t*, updates his expected sojourn time, M_{ijt+1}, for this facility, using W_{ijt}. Additionally, according to the CA model, each agent interacts with two neighbours, who provide him with information regarding their latest experience. Then, agent A_i uses this information to update his expectation with regard to the facility chosen by his quickest neighbour. With this information in mind, agents update their expected sojourn time for their chosen facility and that of their best performing neighbour using an exponentially weighted average with weight α, which is assumed to be constant. This weight is also known as the coefficient of expectations (Nerlove 1958) or smoothing parameter (Gardner Jr. 2006). Given the assumption of homogenous agents, α is the same for all agents. So, the updating process of the memory of agent M_{ijt+1} can be expressed by:

$$M_{ijt+1} = \alpha * M_{ijt} + (1 - \alpha) * W_{ijt} \tag{2.4}$$

where M_{ijt} is the previous value of the memory. See Delgado et al. (2011a) for more technical details about the estimation of M_{ijt}.

2.4.2 Estimating the Uncertainty Measure for the Agents' Expected Sojourn Times

The measure of uncertainty, σ_{ijt}, is modelled using the error in the estimation of M_{ijt}. According to Newbold (1988), if M_{ijt} is a smoothing estimation of W_{ijt}, the error in such an estimate will be:

$$e_{ijt} = W_{ijt} - M_{ijt-1} \tag{2.5}$$

and the cumulative squared error at time *t* can be estimated by the sum of the squared errors:

$$SS_{ij} = \sum_{t=2}^{tsim} e_{ijt}^2 = \sum_{t=2}^{tsim} (W_{ijt} - M_{ijt-1})^2 \tag{2.6}$$

where *tsim* is the simulation time. However, in this way, all the observations are given the same weight; a more realistic approach is to give a larger weight to more recent errors. So in this context, the uncertainty, σ_{ijt}, may be estimated using the concept of volatility forecasting which is calculated by means of the smoothing variance (e.g. Taylor 2006). As the variance is unobservable, we can apply exponential smoothing in order to estimate this variance using the squared residuals (Taylor 2004). Thus, the smoothed variance, σ_{ijt+1}^2, will be expressed as a weighted average of the previous estimate, σ_{ijt}^2,, and the new observation of the squared error e_{ijt}^2,. Thus, agents update the variance, σ_{ijt}^2,, as follows:

$$\sigma_{ijt+1}^2 = \gamma * \sigma_{ijt}^2 + (1-\gamma) * (W_{ijt} - M_{ijt-1})^2 \tag{2.7}$$

where γ is the smoothing parameter (Taylor 2004). The volatility is then measured by the standard deviation σ_{ijt}.

2.4.3 Estimating the Expected Upper Bound of the Sojourn Time

Once the agents compute their expected sojourn time and the uncertainty of this estimate, they consider these values to assess the expected upper bound of the sojourn time, i.e. the estimate of the maximum sojourn time they think they could experience given their expectations and how risk-averse they are. Given the aforementioned "risk-aversion factor", *R*, which is identical for all agents, the expected upper bound of the sojourn time, U_{ijt}, of agent A_i at facility Q_j in period *t* can be written as follows:

$$U_{ijt} = M_{ijt} + R * \sigma_{ijt} \tag{2.8}$$

The agents' uncertainty, σ_{ijt}, is assumed to be the volatility of their expectations and the risk-aversion factor, *R*, may be considered as how sensitive agents are to this volatility. The larger *R* is, the more risk-averse the agents are.

In order to decide which facility to join, agents must update U_{ijt} each period. According to the decision rule explained in the CA Model, they always patronise the facility with the lowest value of U_{ijt}, i.e. agents update their state by choosing the queue with the lowest expected upper bound of the sojourn time. In the rare case where two or more queues have the same minimal expected upper bound, agents choose between these facilities, giving first preference to their previously chosen queue and second choice to the one previously used by their best performing neighbour.

2.4.4 Average Sojourn Time in a Transient Period

In this chapter, we consider a queueing system whose arrival rates may temporarily exceed the service rates. Hence, we need a measure for the average sojourn time that enables us to study the system behaviour in a transient state, rather than in steady state

This measure is proposed and explained in Sankaranarayanan et al. (2011). They consider a congestion measure which satisfies the well-known Little’s Law and the steady state equations (Gross and Harris 1998), while remaining well-defined when $\rho \geq 1$ (transient analysis). This measure is:

$$W_{jt} = \frac{\lambda_{jt}}{\mu^2} + \frac{1}{\mu} \qquad (2.9)$$

where μ is the service rate for all facilities and λ_{jt} the number of agents arriving at Q_j at time *t*. For more details about the formulation of this measure, see Sankaranarayanan et al. (2011) and Delgado et al. (2011a). We adopt this measure throughout this chapter.

2.5 Characterisation of Customers

Customers (referred to as agents in this chapter according to the terminology of CA models) can be classified according to the values of their parameters. We characterise the different customer types as follows:

- Customers are considered risk-neutral when $R = 0$. We can also call them customers who ignore the uncertainty. Otherwise, if $R > 0$, we say that customers account for uncertainty and they therefore have a certain level of risk-aversion. We assume that customers have a low risk-aversion when $0 < R < 0.4$. For $0.4 \leq R \leq 1.2$, we say that their risk-aversion is intermediate, while for $R>1.2$, we say that customers are very risk-averse.
- Depending on the coefficients of expectations (α and γ), customers can be called conservative or reactive regarding the new information:
  - For values of $\alpha > 0.5$, we shall say that customers are conservative with regard to the expected sojourn times, i.e. they give little weight to the new information regarding their or their neighbours’ most recent experience. Alternatively, if $\alpha < 0.5$, we shall call them reactive customers regarding the expected sojourn times, since they attach more importance to the new information than to the past.
  - A similar reasoning applies to γ: Customers are considered to be either conservative or reactive regarding the use of new information to estimate the variability of their expectations when, respectively, $\gamma > 0.5$ or $\gamma < 0.5$.

2.6 Simulation Results and Discussion

The agents of a CA model are endowed with memory (Wolfram 1994). This feature enables us to use this framework to investigate the problem we address here. We model the agents' memory using adaptive expectations as described above. As the system behaviour depends on the initial values of memory assigned to the agents, i.e. the evolution of the system is path dependent, our model cannot be solved analytically. Hence, we use simulation to understand the system behaviour. For its implementation, we use MATLAB 7.9, a numerical computing environment used in engineering and science (The MathWorks 2009)

The CA model is configured with 120 agents (i.e. the number of cells n in the one-dimensional discrete lattice) and 3 facilities (i.e. the number of states m each cell may take). As we simulate a system with just three facilities, considering a neighbourhood size larger than 1 would be to assume that customers could often have full information. We therefore use a neighbourhood size (K) equal to 1. The service rate is the same for all facilities and equals 5 agents per unit of time. Each agent is allocated an initial memory for the expected sojourn time for each facility. These memories are distributed randomly around the optimal average sojourn time. This setup of the system is appropriate to observe the phenomena with which we are concerned.

In order to study the impact of including uncertainty in the agents' decisions, we divide the simulation results in three parts: the first part illustrates the simulation of a typical example in which we compare the resulting collective behaviour of customers when they include or ignore uncertainty. We refer to this example as the base case. Its results are reported in Figure 2.1 to Figure 2.5. We then study the distribution of average sojourn time for different values of the risk-aversion factor, R, while keeping the smoothing parameters constant. This distribution is illustrated in Figure 2.6. Next, we perform a sensitivity analysis with regard to the smoothing parameters α and γ. Figure 2.7 shows the outcomes of this sensitivity analysis.

Table 2.1 summarises the parameters used to configure the system for the base case. Numerous simulations of the model enable us to say that 100 periods is enough time to observe and analyse a typical example of how the system behaves during and after the transient period. Thus, one run of the model in this first part of the analysis is carried out over

100 simulation periods. This setup is used to simulate one case where customers include uncertainty ($R = 0.5$) and another one where they ignore it ($R = 0$).

Parameter	*Value*	*Description*
M	*3*	Number of service facilities
N	*120*	Population size
μ	*5*	Service rate
α	*0.3*	Weight to memory when updating the expected sojourn time
γ	*0.7*	Weight to memory when updating the estimated variance of the expected sojourn time
Tsim	*100*	Simulation time

Table 2.1 Parameter values used for the simulation run.

Figure 2.1 and Figure 2.2 show the evolution of the agents' choices of service facility over 100 periods for the same initial values of expected sojourn times, but assuming different risk attitudes. Figure 2.1 involves risk-neutral agents (R = 0, i.e. uncertainty ignored), while risk-averse agents (R = 0.5, i.e. uncertainty included) are concerned in Figure 2.2. The horizontal axis depicts the time, while the 120 agents are represented on the vertical axis from top to bottom. Colours indicate the state of each agent each period. Black indicates facility 1, grey facility 2 and white facility 3.

Both figures exhibit an initial transient period. This period is longer when agents consider uncertainty (R = 0.5) in their decisions, around 34 periods (see Figure 2.2) compared to the case where uncertainty is ignored (R = 0), around 15 periods (see Figure 2.1). This result is undesirable for risk-averse customers who would prefer a system that stabilises quickly. The stabilisation of the system depends on how fast the elements considered by the customers to choose a facility converge to a target. Risk-averse customers base their decision on the expected upper bound of sojourn time (U_{ijt}), which is formed by their expected sojourn time (M_{ijt}) and an estimate of their uncertainty, while risk-neutral customers only consider the expected sojourn time. The target for the expected sojourn time (M_{ijt}) is the average sojourn time at each facility (W_{ijt}) (see Equation 2.4), while the target for the estimated uncertainty (M_{ijt} - W_{ijt}) is zero. Given that the estimate of uncertainty depends on the expected sojourn time, the former converges slower than the latter. As a system with risk-averse customers requires that both elements converge to the respective targets in order to stabilise, this process takes more time than for a system with risk-neutral customers, which only requires the convergence of the customers' expected sojourn time. Consequently, due to their cautious behaviour, risk-averse customers achieve the opposite of what they wish.

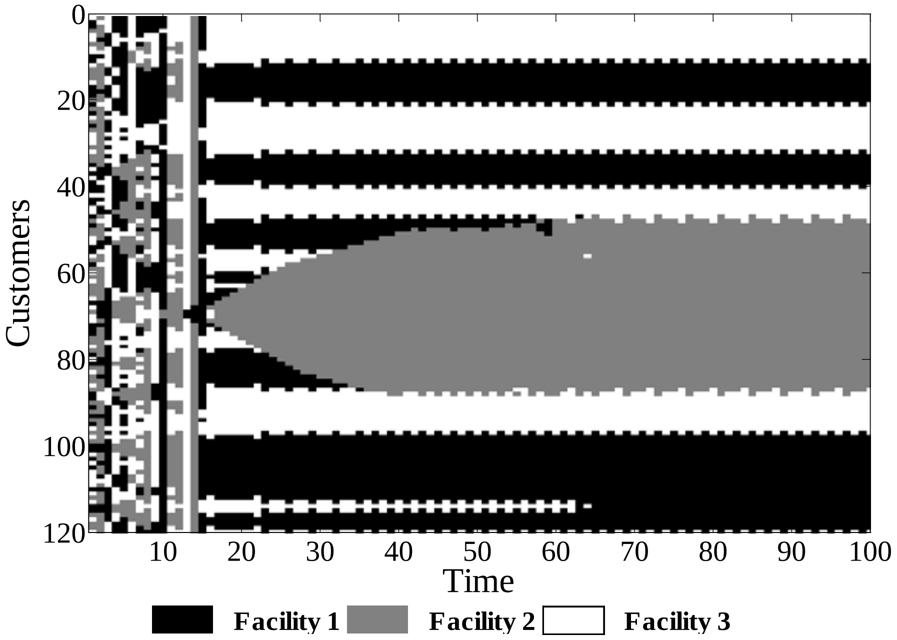


Figure 2.1 Spatial-temporal behavioural evolution of risk-neutral agents' choice of service facility with α= 0.3; γ = 0.7 and R = 0 (uncertainty ignored)

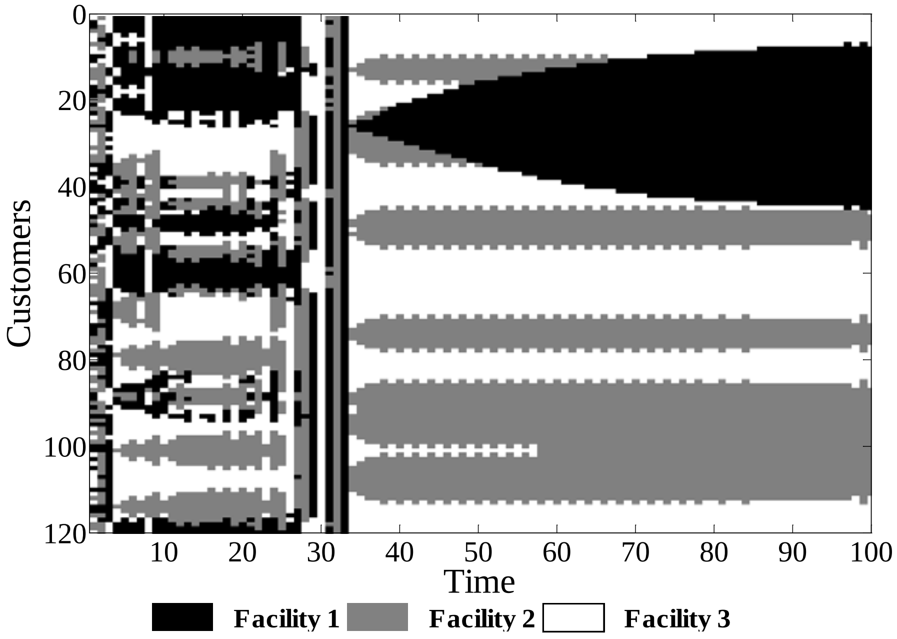


Figure 2.2 Spatial-temporal behavioural evolution of risk-averse agents' choice of service facility with α= 0.3; γ = 0.7 and R = 0.5 (uncertainty included)

The length of the transient period also varies depending on the randomly allocated initial expected sojourn times (Delgado et al. 2011a). During this period, agents explore all facilities in order to capture information and try to learn from the system behaviour. The weight they give to this information leads to herding behaviour: agents tend to imitate their best performing neighbours and at the end of the transient period one or two facilities tend to be crowded, as shown in Figure 2.1 (between periods 9 and 15) and Figure 2.2 (between periods 27 and 34). As more weight is given to new information than to memory ($\alpha < 0.5$), agents react to the bad experiences by changing facility at the next period. For instance, in Figure 2.1 facility 1 is crowded at time 10, implying that no agents choose this facility in the next two periods. A similar situation occurs in Figure 2.2 at time 30 when all agents join facility 3 resulting in that this facility not being used for the next three periods.

After the transient period, a collective behaviour pattern emerges in Figure 2.1 and Figure 2.2. We explain this pattern in two stages. The first stage portrays a diffusion phenomenon in which some of the agents, who previously experienced a very bad experience at a facility, start to go back to this one. Due to the sharing of information between neighbours, the number of returning customers increases gradually.

Figure 2.1 and Figure 2.2 illustrate two different situations from which this phenomenon can stem as discussed below. In the second stage, stable groups of customers loyal to a facility emerge. In Figure 2.1, which concerns risk-neutral agents ($R = 0$), the diffusion phenomenon involves facility 2 (grey) and appears in period 16, when two agents (68 and 71) decide to try once again this facility after a bad experience at facility 1 (black). In that period, these two agents are the only ones to join facility 2, since all other agents consider this facility as a very poor choice. As a result, these two agents experience an extremely low sojourn time, the news spreads and other agents return. The cause of this phenomenon lies in period 14, when facility 2 was the most patronised. However, four agents did not use that facility in that period (including agents 68 and 71). Hence, the bad performance at this facility in period 14 did not affect their memory, while that of the other agents increased a lot. At time 15 these four agents stay at facility 1, which is the most visited. Consequently, the expectations of agents 68 and 71 regarding this facility soar and these agents decide to move to facility 2 in period 16, where they have a good experience as mentioned above. Over the next periods, they share this experience with their neighbours, who start coming back to facility 2. Then, the number of agents patronising this facility gradually increases and a group of loyal agents to this facility starts to emerge. This phenomenon persists until some agents on the spatial borders of the group are disappointed and decide to try the facility which their best performing neighbour used one period early. On the lower border, this occurs in period 40 when agent 88 is the last one to join facility 2. On the other side, the diffusion continues, but more slowly, until period 66. Agent 47 is the last one to join this facility. After period 42, the last agents joining facility 2 continue to alternate irregularly between this facility and that of their best performing neighbour.

As for Figure 2.2 which involves risk-averse agents (R = 0.5), the diffusion phenomenon arises in period 34 and is owed to one sole agent (26), who stays at facility 1 (black) despite his bad experience there one period earlier (period 33). The remaining agents consider this facility as a bad choice and they therefore move to facilities 2 (grey) and 3 (white). Thus, agent 26 experiences the lowest sojourn time in the system in period 34. This event considerably improves his expectation regarding this facility. Moreover, his neighbours hear about this experience and decide to try again this facility in the next period (35). Later, the information spreads to further neighbours resulting in more agents coming back to this facility bit by bit.

The collective pattern then continues with small groups of agents tending to stay at the same facility over time, resulting in a certain degree of stability for the system. We call these agents loyal. However, the agents located on the borders between these groups keep shifting between two facilities in a stable way. For example, in Figure 2.1, agents 47 and 88 switch between facilities 2 (grey) and 3 (white) following the sequence {2-3-3-3-2-3-3-3-2…}, while agents 48 and 87 move between the same facilities with the sequence {3-2-2-2-3-2-2-2-3…}. Comparatively, this form of stability cannot be identified in figure 2, since at the end of the simulation period (100), no regular pattern is yet reached. Such a system requires more time to stabilise.

In Figure 2.1, agent 56 illustrates an odd case in period 64: This agent decides to leave facility 2 (grey) in period 64, after having patronised it for a long period, because of his expectation regarding this facility exceeds that of facility 3 (white). This occurs because after the transient period, this agent patronises for some time facility 3 resulting in his expectation regarding this facility improving significantly (see Figure 2.3). Additionally, when he leaves this facility in period 30, this is the last time he updates his memory regarding this facility owing to his quickest neighbour never using it again either. Thus, his expectation regarding facility 3 remains low and close to that of facility 2, as shown in Figure 2.3. Hence, when the number of agents using facility 2 increases in period 63, the expectation of agent 56 regarding this facility exceeds that of facility 3. He decides therefore to try facility 3 in period 64. However, he experiences a bad sojourn time (note the increase of his expected sojourn time for facility 3), which encourages him to come back to facility 2 and stay there for the remainder of the simulation.

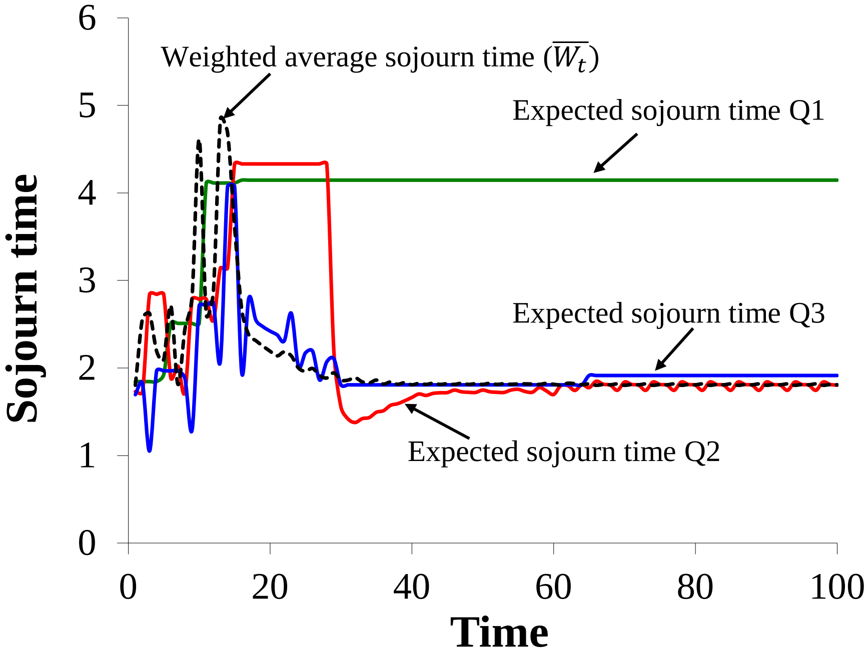


Figure 2.3 Expected sojourn times of agent 56 when he ignores uncertainty (*R* = 0), i.e. a risk-neutral agent.

Figure 2.4 and Figure 2.5 show the evolution of the weighted average sojourn time of the system and the minimum and maximum sojourn times experienced by customers. the weighted average sojourn time is computed using the following equation:

$$\overline{W}_t = \frac{\sum_{j=1}^{Nqueue} W_{jt} * \lambda_{jt}}{N} \tag{2.10}$$

where λ_{jt} is the number of customers patronising facility j at time t and W_{jt} is the average sojourn time these customers experience at this time at this facility.

In Figure 2.4 and Figure 2.5 we can observe that in those periods when most agents patronise the same facility the maximum, minimum and average sojourn times experienced by the agents reach extreme values. For instance, in the example where agents ignore uncertainty

(Figure 2.1 and Figure 2.4) most agents choose facility 3 in period 13, two agents facility 1 and no agents facility 2. Hence, the two agents at facility 1 experience the lowest sojourn time ($W_{1,13}$ = 0.28), while the agents at facility 3 experience a very high sojourn time ($W_{3,13}$ = 4.92). A similar case occurs in Figure 2.2 and Figure 2.5, in period 34 when a single agent joins facility 1 and experience the lowest sojourn time ($W_{1,34}$ = 0.24).

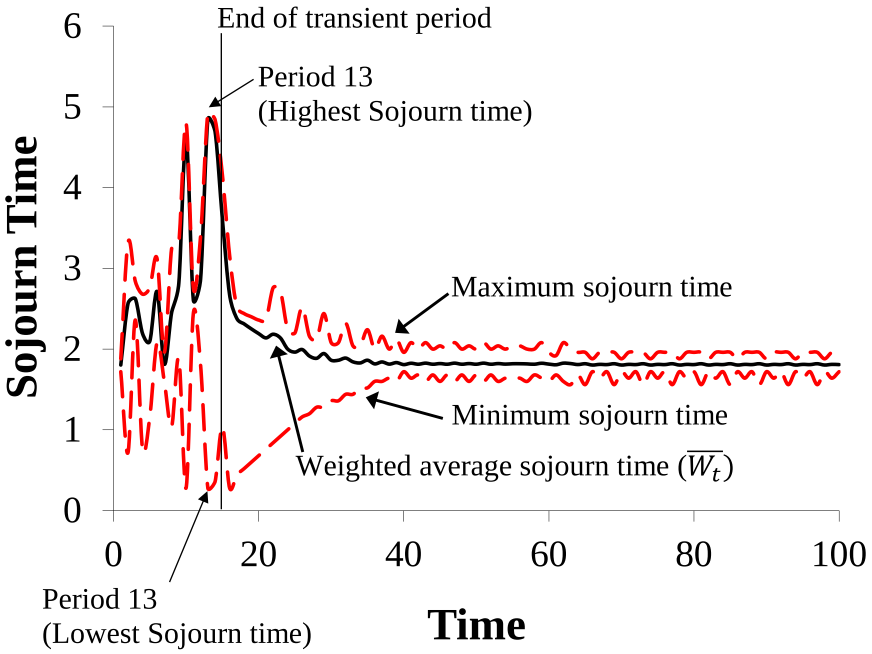


Figure 2.4 Average, maximum and minimum sojourn time for a system configured with risk-neutral agents (R =0) and coefficients of expectations: α= 0.3; γ = 0.7

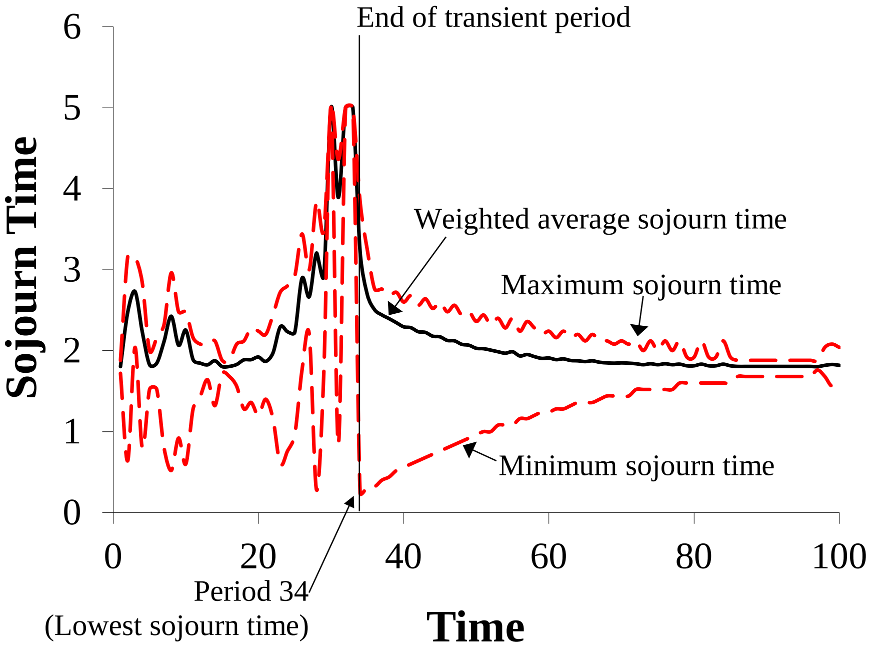


Figure 2.5 Average, maximum and minimum sojourn time for a system configured with risk-averse agents (R =0.5) and coefficients of expectations: α= 0.3; γ = 0.7

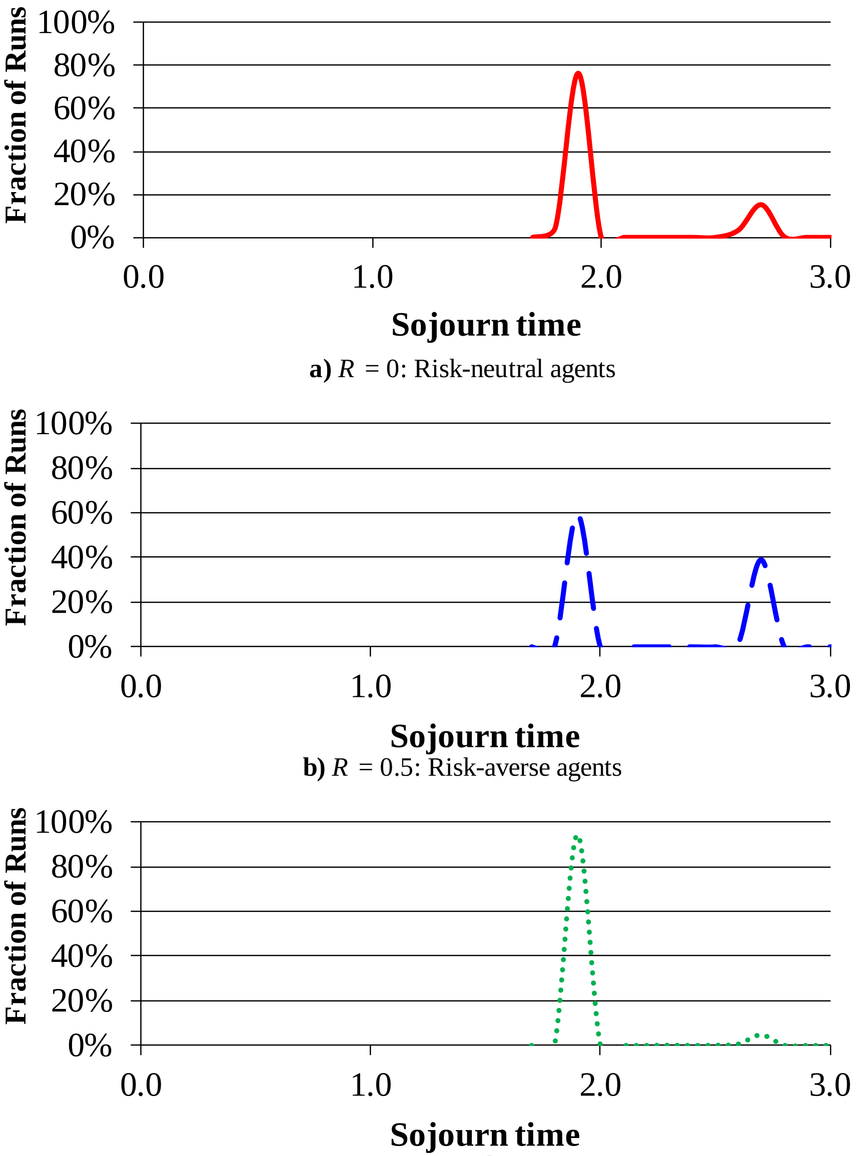
Figure 2.4 and Figure 2.5 also illustrate the general observation that a system in which agents are risk-neutral converges faster to an almost-stable average sojourn time than a system with risk-averse agents.

The discussion provided so far has been based on two examples, but the patterns of behaviour we discussed are typical of what we have observed over several thousands of simulations. We have also simulated the cases with agents who are even more risk-averse. The observed behaviour is similar, but with a longer transient period: stable patterns start to emerge after 500 or more periods.

Figure 2.6 shows the distributions of the weighted average sojourn time for 1,000 simulations of a system configured according to the parameter values of Table 2.1 with R equals to 0 (uncertainty ignored), 0.5 (an intermediate degree of risk-aversion) and 1.5 (very high risk-aversion). Each simulation was run for 1500 periods with different initial expected sojourn time allocated to the agents randomly. The weighted average sojourn time was calculated based on the last 500 periods of each run.

Figure 2.6 illustrates the dependence of the sojourn time on the allocated initial expected sojourn times and enables us to identify the probability of a facility being ignored by the agents in steady state. Figure 2.6 a) depicts the distribution of weighted average sojourn times for the system where agents ignore uncertainty, while in Figure 2.6 b) and 2.6 c) agents incorporate uncertainty into their decisions. The first peak of each distribution represents the proportion of cases where the three facilities are used in steady state. When R = 0, around 81% of the cases reach a weighted average sojourn time in steady-state between 1.80 and 1.90 time units, compared to 57% of the cases for R = 0.5 and 94% for R = 1.5. The remaining cases (19% for R = 0, 43% for R = 0.5 and 6% for R = 1.5) fall in the interval [2.6, 2.8]. In these cases one facility is ignored when the system has reached a steady state, i.e. all the agents are clustered in only 2 of the 3 facilities.

Given the previous analysis, we can conclude that the more risk-averse the customers, the longer the transient period the system exhibits. Moreover, Figure 2.6 shows that there is a non-monotonic relationship between the degree of risk-aversion and system performance; customers with an intermediate degree of risk-aversion typically are more likely to ignore a facility than those who are very risk-averse or risk-neutral. Consequently, very risk-averse customers and risk-neutral customers achieve lower average sojourn times.



a) *R* = 0: Risk-neutral agents

b) *R* = 0.5: Risk-averse agents

c) *R* = 1.5: Very risk-averse agents

Figure 2.6 Distributions of weighted average sojourn time for 1,000 simulation with different initial conditions for α= 0.3; γ = 0.7

2.7 Sensitivity Analysis

In this section, we present a sensitivity analysis with respect to the risk-aversion parameter (*R*) and the expectation coefficients (α, γ).The model is configured with the same settings as in the previous section (see Table 2.1), but varying the parameters α and γ. It was tested for different initial expected sojourn time (M_{ij0}) allocated to the agents randomly. This allows us to identify that after 1,000 periods the system starts to exhibit a certain stability, i.e. the variance of the sojourn time is less than 10% and a collective behaviour pattern of agents over time is easily identified and characterised. Hence, we run the model for 1500 periods and the weighted average sojourn time is computed for the last 500 periods.

We adopt the definitions of customer types given in Section 2.5. Figure 2.7 illustrates how the weighted average sojourn time varies depending on the value of these parameters. Each graph in Figure 2.7 shows, for a given expectation coefficient of the expected sojourn time (α), the weighted average sojourn time of the system as a function of the expectation coefficient of the variance (γ) and the risk-aversion parameter (R). The curves in each graph represent the weighted average sojourn time for different values of the expectation coefficient of uncertainty (γ) depending on the risk-aversion level of customers (R, horizontal axis). These results are based on 1,000 simulations of the model for each combination of the parameters α, γ, and R. We simulate the model for each combination of parameters using the same random seeds.

We have mentioned above that the weighted average sojourn time is higher than 1.8 and that those values falling in the interval [2.6, 2.8] indicate that customers have ignored a facility. Thus, the higher the weighted average sojourn times exhibited in figure 2.7, the higher the probability that in steady-state a facility is being ignored.

A facility j is ignored when for all customers the estimated upper bound of the sojourn time (U_{ijt}) at this facility is much higher than that of the other two facilities. When this occurs, customers will not receive new information to update U_{ijt} over the next periods. Moreover, when the customers' estimated upper bound of sojourn time for at least one of the other two facilities is close to the target (i.e. the average sojourn time), the facility currently being ignored will never be used again. Consequently, customers will either stay at one of the two remaining facilities or alternate between these.

The first graph of Figure 2.7 ($\alpha = 0.1$) illustrates the case where customers give significantly more weight to new information than to the past when updating their expectations of sojourn times. In this case, customers with a very low level of risk-aversion (i.e. R close to 0) perform poorly, i.e. the system reaches high average sojourn times. As α is very small, the expected average sojourn time (i.e. the memory M_{ijt}) is strongly affected by new information. Thus, extreme experiences at a facility (i.e. very high or very low sojourn times) will significantly impact the customers' expected sojourn time as well as the squared error of this expectation. However, since R is also very small, whatever the value of γ, uncertainty plays a negligible role when estimating the upper bound of sojourn time. Consequently, when customers experience extreme sojourn times, their expected upper bound for the next period will be well above the average sojourn time and therefore the probability of ignoring a facility increases.

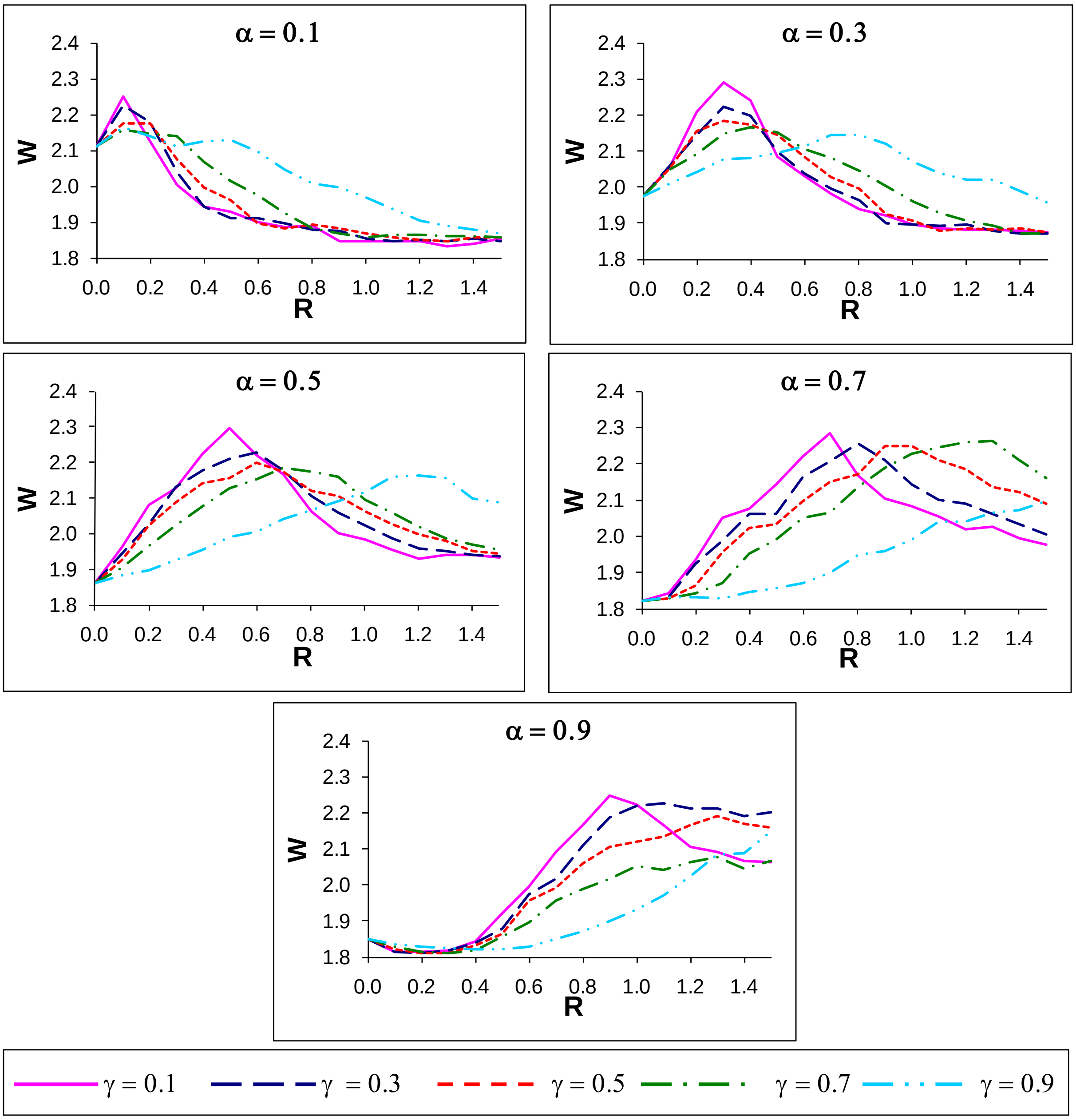


Figure 2.7 The weighted average sojourn time of the system as a function of the coefficients of expectations (α , γ) and the risk-aversion parameter (R).

An extreme experience, whether positive or negative, will have a strong impact on the expected sojourn time: a decrease in case of a positive experience, and an increase in case of a negative experience. But, in both instances, the estimate of the uncertainty will increase as the experienced value will be very different from the customer's expectation, thus significantly increasing the second component of the estimate of the upper bound given the large value of R. Thus, in the case of a positive experience, the decrease in the estimate of the expected sojourn time will be partially, if not fully, offset by the increase in the uncertainty component of the upper bound. On the contrary, in the case of a negative experience, the increase in the uncertainty will reinforce the increase in the expected sojourn time when estimating the upper bound. Consequently, larger values of R lead to larger values of the estimated upper bound, implying that customers will be more inclined to test other facilities, thus increasing the length of the transition period.

As α (i.e. the expectation coefficient to update the expected sojourn time) increases, very risk-averse customers tend to perform increasingly worse (i.e. higher average sojourn times), while risk-neutral customers do better. Note also that, whatever the value of γ , the value of R for which customers are most likely to ignore a facility (i.e. very high weighted average sojourn times), increases in α. However, this does not imply that when α is very high, very risk-averse customers performs very poorly. On the contrary, we can observe in Figure 2.7 that when customers are very conservative (α = 0.9) with regard to their expected sojourn time and less conservative with regard to the variance ($\gamma < 0.9$), the average sojourn time starts to decrease in R once this parameter exceeds a certain threshold.

Customers who are very conservative regarding both the expected sojourn time and the variance ($\alpha \geq 0.7$ and $\gamma \geq 0.9$) are a special case. Note that in these situations, the relationship between R and the system performance is monotonic: the more risk-averse the customers, the more likely they are to achieve higher average sojourn times and to ignore a facility. Given that α and γ are high, the new information does not matter very much. Thus, extreme experiences at a facility (i.e. very high or very low sojourn times) have a very weak impact on both the customers' expected average sojourn time (i.e. the memory) and their estimate of the variance. Hence, as the risk-aversion level (R) increases, it plays a more important role when estimating the expected upper-bound than the other two parameters. Moreover, when R is very high, this parameter dominates, i.e. the estimated upper bound of the sojourn time is barely affected by the expected sojourn time. Thus, the estimated upper bound of the sojourn time increases and tends to be well above the average sojourn time. As the weight given to the new information is very small, the expected upper bound will converge very slowly to the average sojourn time. Consequently the probability of ignoring a facility increases and the system achieves, on average, higher weighted average sojourn times than when customers have a lower risk-aversion level.

Regarding the other combinations of parameters α and γ, the relationship between R and the system performance is non-monotonic, as shown in Figures 2.6 and 2.7. For instance, customers who update their perception of the variance more quickly ($\gamma = 0.1$) systematically achieve the highest average sojourn times (i.e. the peak of the pink lines in Figure 2.7) when R equals α. In such situations, customers are more likely to ignore a facility forever. Note that when R exceeds 1 (the maximum value which α can take), the weighted average sojourn time decreases in R. The analysis of this non-monotonic relationship is very complex and would require studying in more detail the interaction between the three parameters (α, γ and R).

According to Figure 2.7, we can describe the performance of customers depending of the weight they give to their memory (α). For $\alpha \leq 0.3$ (i.e. reactive customers according to the customer types of Section 4), very risk-averse customers ($R > 1.2$) perform better than risk-neutral customers do (R =0); while for $\alpha \geq 5$ (i.e. conservative customers), risk-neutral customers do better than risk-averse customers ($R > 0$). In the extreme case where customers are very conservative regarding their expectations of sojourn time ($\alpha = 0.9$) the average sojourn time achieved by customers with low risk-aversion is close to the Nash Equilibrium, which is equal to 1.8 periods. The system achieves such an equilibrium only when each period an equal number of customers patronise each of the three facilities. The Nash equilibrium is the optimal behaviour the system could achieve. However, the best performance which the system achieves yields a sojourn time a little higher than the Nash equilibrium. This performance occurs when the system is set up with customers characterised as:

- Very conservative when updating their expectations of sojourn time (i.e. $\alpha = 0.9$),
- Rather conservative when updating their expectations of variance (i.e. $\gamma \geq 0.5$), and
- With lower risk-aversion levels (i.e. $0.1 \leq R \leq 0.3$).

Next, let us consider the impact of the coefficient of expectations to update the estimate of the variance (γ). Conservative customers (γ high) with an intermediate level of risk-aversion perform well as long as the parameter α remains above 0.3. This is particularly the case when the parameters α and γ are very high (i.e. α and $\gamma = 0.9$, which means that customers are reluctant to consider new information to update their expectations). Very risk-averse customers experience, on average, lower sojourn times when they consider the new information as very important ($\gamma < 0.5$) to update the variance, except for the cases where they are very conservative regarding their expectations of sojourn time ($\alpha = 0.9$). The lower γ, the spikier the behaviour of the average sojourn time is as a function of risk-aversion (R). That means, the average sojourn time is more sensitive to small changes.

2.8 Conclusions and Future Work

We have modelled a service system with interacting customers who must decide each period which facility to join for service. For this system, we have studied the impact of accounting or not for uncertainty in the customers' decisions on the collective behaviour of customers and on the weighted average sojourn time of the system. A one-dimensional CA model has been used to describe how customers interact with their neighbours and share information regarding their experiences. Risk-neutral customers base their decision on the expected sojourn time (i.e. they ignore uncertainty), while risk-averse customers estimate an upper bound for the different sojourn times. Risk-averse customers use their experience and that of their neighbours to compute this upper bound, by using their expected sojourn times, their estimate of the uncertainty concerning this expectation and a risk-aversion parameter. They estimate their expected sojourn times and their uncertainty by applying adaptive expectations.

The model has been simulated for different combinations of parameters that characterise the customer types. These simulations show that the more risk-averse the customers, the longer the transient period exhibited by the system and the slower the convergence to an almost-stable weighted average sojourn time. This outcome is undesirable for risk-averse customers who would prefer a system that stabilises quickly, but occurs because the more risk-averse customers are, the slower the estimated upper bound the sojourn time converges to the average sojourn time.

Very risk-averse and risk-neutral customers are less likely to permanently ignore (i.e. never use it again) a facility in the long run than those who have an intermediate degree of risk-aversion. Permanently ignoring a facility is unrealistic if it remains in operation. In such a case, sooner or later a customer will return, whether by accident (random choice) or by curiosity (he has not been there for a long time) and experience a very low sojourn time. He will share this information with his neighbours, thus starting a diffusion phenomenon through which the number of returning customers will increase gradually.

Systems where customers are either close to risk-neutral or strongly risk-averse perform better than those who have an intermediate level of risk aversion. Very risk-averse customers experience low sojourn times (i.e. good performance) when they give significant weight to the most recent experience to update their memory regarding expected sojourn time (i.e. α is small). Moreover, they achieve their best performance when they update their perception of the variance more slowly. If these very risk-averse customers are reluctant to take into account new information to update their expectations of sojourn time (α large), they will experience higher sojourn times. Risk-neutral customers and those with low risk-aversion achieve their best performance when they give little weight to the most recent experience when updating their expected sojourn times.

As far as customers with an intermediate risk-aversion level are concerned, they perform better when they use either low weights to update their two expectations (i.e. expected sojourn time (α) and the variance (γ)) or high to both.

The optimal choice of updating parameters depends on the customers' risk attitude. Consequently, future research will focus on studying the impact of the different behavioural parameters (α, γ, R). This will include allowing for different levels of customers’ reactivity depending on the source of the information, i.e. giving different weights to own experience and information received from neighbours in the memory updating processes (expected sojourn times and variance). The next step will be to assume heterogeneous customers. In particular, we will consider customers with different degrees of risk-aversion (R) and/or different levels of reactivity (α, γ). Another interesting aspect would be to focus on the service capacity. For example, assessing the collective behaviour when the facilities have different service capacity or, more interestingly, assuming that managers are able to adjust the service capacity of the facilities depending on the customers’ behaviour (i.e. endogenous service rates).

3 Queueing Systems with Interacting Customers and Service Providers

ABSTRACT

We address a service facility problem with reactive customers and managers. This problem is modelled as a deterministic queueing system in which customers must routinely choose a facility for service and managers are able to adjust the service capacity. Customers cannot observe queues before choosing a facility. However, they interact with their neighbours and share information regarding their previous experience. Both managers and customers base their decisions on their perceptions about the system. They are endowed with computational memory to update their perceptions using an adaptive expectations model. Customers use their previous experience and that of their neighbours to update their perceptions about the average sojourn time, while managers form their perceptions based on the queue length. We use cellular automata to model the interaction between customers and managers. We perform a simulation to assess the way the customers' and managers' decisions evolve and affect the system behaviour. Our results show that the system we study exhibits a certain degree of path-dependence. The main conclusion is that the more conservative managers tend to achieve a larger market share.

KEYWORDS: Queueing problems, endogenous service and arrival rates, cellular automata (CA), adaptive expectations, path-dependence.

3.1 Introduction

The preceding chapter addressed the problem of a service facility system in which customers face the situation each period of having to choose between m facilities for service. For instances, car owners who annually or biannually (depending on the country) must choose a garage for emission tests, students or employees who daily look for a restaurant for lunch, and so forth. The model built in the previous chapter assumed that all facilities had the same service capacity which did not change over time, i.e. exogenous service rates. Relaxing this assumption by enabling managers to make decisions regarding the facilities' service capacity, we attempt to go onwards in studying the behavioural aspects of queueing problems with repeat customers. In many real-life service systems, managers are able to adjust their service capacity depending on the customer behaviour. They look to retain their current customers and attract new ones to their facilities in order to increase their profits. The new customers can either be customers currently patronising another facility or potential customers who wish to enter the system (e.g. new car owners and new students). In this chapter, we focus only on existing customers, while new potential customers will be tackled in the next chapter.

As managers value their customers because they increase the value of the firm customers value their time. "Time is money", such as the adage says. Whatever the service customers require, waiting for service represents a waste of time for them which affects their utility. This impact is even stronger when customers repeatedly patronise a facility for service. When customers perceive that their utility is being affected, they look for another service provider who maximises their utility. The complexity of the relationship between managers and customers increases in real life when many service facilities compete to render a same service. In this case, the managers' actions will affect their future decisions, those of the customers as well as those of the rival managers.

A very broad range of studies has addressed the behaviour of customers and managers in queueing problems. Nevertheless, this literature is scattered and not well-organised. The literature related to customer behaviour has been broadly discussed in the previous chapter and in the introduction of this thesis. The research on customer behaviour in queueing systems has been mainly tackled by marketing researchers, who study the relationships among waiting times, customer satisfaction and service quality in service facilities (Davis and Heineke 1998; Hui and Tse 1996; Taylor 1994). These studies attempt to understand the influence of waiting time on customer satisfaction, customer loyalty and service quality

(Bielen and Demoulin 2007; Law, et al. 2004). Their aim is to endow managers with information about customers’ attitudes to enable them to redesign their service facility accordingly. For a review of the literature, see in Bielen and Demoulin (2007) and Gallay (2010).

Concerning the managers’ decisions, most studies in the literature have focused on analysing policies of optimal pricing and capacity decisions to control problems associated with congestion in service facility systems. P. Naor (1969) is the seminal paper on this subject. He formalised the insight originally formulated by W. A. Leeman (1964) and then discussed by T. L. Saaty (1965) and W. A. Leeman (1965). These authors suggest using pricing to help reduce queues in many service systems. Naor’s model was subsequently generalised by Yechiali (1971), Edelson (1971), Edelson and Hilderbrand (1975), Stidham Jr. (1985, 1992), Mendelson and Whang (1990), Dewan and Mendelson (1990); van Ackere (1995), among others. More recently, Sinha, et al. (2010) applied an optimal pricing scheme of surplus capacity to control the joint problem of existing and potential customers who are differentiated according to a pre-specified service quality level.

Although some managers’ strategies effectively consider either the demand or the supply perspective when adjusting their service capacity, optimal strategies should incorporate the perspective of the two conflicting parts of the system (Pullman and Thompson 2002). Our research is motivated by the logic behind this assertion and the complexity of the interaction between the decisions of customers and managers in service facility systems.

Consequently, our modelling approach considers a service facility system where competing facilities render a service which customers require routinely. Each facility has its own queue and manager. Queues are assumed to be invisible to the customers. We assume that customer interact with their neighbours and share information about their most recent experience. They use their experience and that of their best performing neighbour to update their expectations of their previously chosen queue and the one used by their quickest neighbour. Then, based on their expectations, customers choose a facility for the next time. Managers take their decision to adjust service capacity on the basis of their desired service capacity which they determine based on their perception of the queue length at their facility and a market reference sojourn time. This market reference is a benchmark whereby managers compete with each other to attract more customers to their facilities. In other words how managers perform compared to this benchmark is a competitiveness index of the facilities.

In order to study this complex problem we propose an idealised queueing model with reactive and adaptive customers and managers in which the decisions of both types of agents are interdependent. This model is built using a CA-based framework. As in the previous chapter, the interaction between customers is portrayed in a one-dimensional cell lattice. The main structure of the CA model is similar to that of the model used in Chapter 2 and in Delgado et al. (2011a, 2011b, and 2011d) (See Appendices B, C and E). However, the factors which determine the average sojourn time customer experience at the facilities are different. While in the previous chapter this experience depended only on customers' decisions because the service capacity remained constant (i.e. exogenous service rates), now this experience is also influenced by the service providers' decisions (i.e. endogenous service capacity).

Our results show that the managers' and customers' behaviour is strongly influenced by a path dependence phenomenon. Once historical or random events determine a particular path, agents may become locked-in regardless of the advantages of the alternatives. In our context, two or three managers can have the same profile, however given the initial conditions, which are randomly allocated to the agents (i.e. customers and managers), their facilities could evolve completely differently. W. B Arthur (1990) explains this phenomenon in the following way: "if a company or a nation in a competitive marketplace gets ahead by "chance", it tends to stay at that level and even increase its lead" (p. 92). Other results indicate that the more conservative a manager is, the larger his market share. Additionally, his facility is less likely to close down. Similarly, the facilities of reactive managers are less likely to remain in operation in the long term. Finally we simulate the model for different values of the benchmark sojourn time which managers use as reference to estimate their desired service capacity. The results of this experiment enable us to say that such a measure acts as an attractor point of the system to which the system behaviour converges.

This chapter is organised as follows. The next section describes the one-dimensional CA model we use to study the agents' behaviour in a multichannel service facility system. First, we explain the differences between this model and that of the previous chapter. Next we deal in turn with the managers' and the customers' decision rules. In the third section, we describe the managers' and customers' profile depending on the model parameters. The next section presents the results. We first analyse some typical behaviour of customers and managers and provide an aggregated view of the system. Then we perform some experiments in which we analyse the influence which the different parameters have on the performance of the facilities. Finally, we present the conclusion and contributions of the chapter.

3.2 The Queueing Model

Consider the queueing system and the CA model explained in the previous chapter. This system consists of a fixed population of N reactive and adaptive customers choosing between m facilities for service each period. In that case the arrival rates (λ) were endogenously determined, while the service rates (μ) were exogenous. In this chapter, we propose to make the service rates endogenous and allow service providers to adjust their service capacity. In this sense, we model a system in which customers are free to choose a facility for service and the service providers adjust their capacity depending on the customers' behaviour. The managers' actions can either encourage or discourage customers to use a certain facility.

It is worth pointing out that this new element concerning the facilities (i.e. variable service rate) does not have a significant impact on the structure of the CA model. The only difference lies in the elements which determine the experience each cell (customer) has when taking (using) the different states (facilities). This experience is the average sojourn time of customers at each facility, which is given by Equation 2.9. This equation defines a congestion measure for customers of a system as a function of the arrival and service rates. Such a function considers that arriving customers can temporarily surpass the service rate in a transient period, but it also satisfies the behavioural characteristics of steady-state. In the previous chapter, we assumed the service rate was fixed (i.e. exogenous). Hence, the average sojourn time of the system only depended on the customers' previous decisions. Now, we assume that service providers are able to adjust the facilities' service capacity (i.e. endogenous service capacity). Consequently, the average sojourn time (i.e. customer's experience) depends on both the customers' and service providers' decisions. In other words, the ability of a facility (state) to be more attractive for customers (cells) than the others depends on the behaviour of all agents in the system.

We use the well-known causal-loop diagram proposed by P. Senge (1990) in order to explain the dynamics between customers and service providers (referred to as managers in the remainder of the chapter) as agents who interact in a service system. We will use the term "agents" throughout the chapter when discussing issues which concern both the managers and the customers. Figure 3.1 portrays the feedback structure between a service facility and its customers. This figure consists of two sectors: the customers' behaviour is to the left and that of the managers to the right. Both sectors are connected by the congestion measure, whose evolution determines the dynamics of the actors in the system. Customers decide which facility to use based on their estimate of sojourn time, while managers decide to adjust service capacity based on their estimate of arrival rates. Examples of this kind of system include a garage where customers take their car for maintenance, and workers or students who daily patronise a restaurant for lunch.

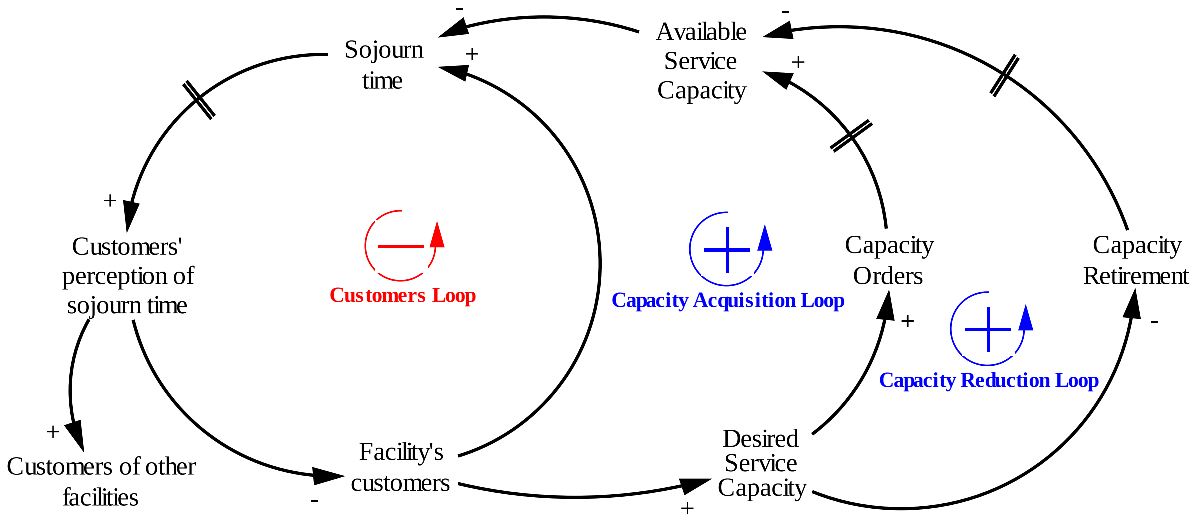


Figure 3.1 Feedback loop structure of the model

3.2.1 Customers' Dynamics

The causal loop in Figure 3.1 highlights the relationship between a service facility and its customers. These dynamics are identical for the *m* facilities. It is worth recalling that we deal with a system portraying a captive market, which means that all customers must patronise one facility for service each period. Customers have the ability to interact with their neighbours in a one-dimensional *K*-neighbourhood. *K* represents the number of neighbours each customer (cell) interacts with on each side. Additionally, we assume that customers have an adaptive behaviour, in the sense that they adjust their memory over time based on their previous experience in the system. Nevertheless, we assume them to be rational when choosing the most convenient facility according to the information they have. This means that they will patronise the facility with the lowest expectation of sojourn time. Thus, customers with a bad perception about a facility can decide to move to another one, as shown in Figure 3.1.

Compared to the previous chapter, we now consider only risk-neutral customers i.e. we assume $R = 0$ in Equation (2.8). Then, customers take their decision based only on their perceptions about the average sojourn time. Additionally, we follow Delgado et al. (2011a) and Sankaranarayanan (2011) and apply different weights for customers to update their memory depending on the sources of the information. In this sense, we denote by α the weight that customers give to their own information when updating their memory regarding their previously chosen facility, and by β the weight for the information provided by their best performing neighbour. In that way, we model the updating process of the customer's memory (M_{ijt+1}) applying an exponentially weighted average of the most recent information (W_{ijt}) and the previous computed expectation (M_{ijt}). Then M_{ijt+1} is given by:

$$M_{ijt+1} = \theta * M_{ijt} + (1 - \theta) * W_{ijt} \tag{3.1}$$

$$\text{s.t. } \theta = \begin{cases} \alpha & \textit{when using own information} \\ \beta & \textit{when using neighbours' infomation} \end{cases}$$

where θ denotes the coefficient of expectations and takes two different values depending on the source of information, as explained above. The logic behind this coefficient is explained in Delgado et al. (2011a) (see Appendix B) and W_{ijt} is computed using Equation 2.9.

To summarise the customers' dynamics: longer (shorter) queues bring about higher (lower) sojourn times and increase (decrease) customers' perceptions. When customers' perception about a certain facility exceeds the expectations they have regarding some other, they decide to switch facility. Otherwise they remain at the same facility. Customers share their experiences with their neighbours who, in turn, update their expectations using the information from their best performing neighbour. Consequently, customers' experience at a certain facility can either encourage or discourage new customers to join this facility. In fact, customers experiencing low sojourn times at one facility can induce their neighbours to use this facility in the next periods. Likewise, customers experiencing high sojourn times can deter their neighbours from joining this facility. The speed at which customers' experience has an impact on their neighbours' memory depends on the weight given to the information provided by the neighbours. Then, the more (less) customers patronise a facility, the longer (shorter) queues are.

3.2.2 Service Providers' Decisions

The two reinforcing loops at the right side of Figure 3.1 (c.f. capacity acquisition and reduction loops) illustrate the managers' dynamics. These dynamics result from the interaction between managers and customers' actions through the state of the system. These two feedback loops describe a very basic type of learning (Sterman 2000), in which managers use the information they have about the system to form expectations, they perceive a gap between the desired and the current state, and they attempt to move the current state toward the desired state. In order to model this feedback process analytically, we endow managers with similar abilities as the customers. In this sense, we assume that managers have a memory and react to customer behaviour by adjusting the service capacity of their facility. Although customers cannot observe the queues before choosing a facility, managers have information about the number of customers arriving at their facilities. They thus use this information to form their perceptions about the future arrival rate, $\hat{\lambda}_{jt}$. The managers' memory also enables them to update their perceptions each period using an adaptive expectation model, as follows:

$$\hat{\lambda}_{jt+1} = \delta * \hat{\lambda}_{jt} + (1 - \delta) * \lambda_{jt} \tag{3.2}$$

where hats indicate the expected queue length, and δ the coefficient of expectations (Nerlove, 1958). δ can be also interpreted as the speed at which managers adjust their perceptions. This parameter follows the same logic as explained above for the customers' parameters (α and β).

Managers use their estimate about the future demand (i.e. arrival rate) to determine the service capacity required to meet their customers' expectations of sojourn time. Managers do not have accurate information regarding these expectations, but they know a reference average sojourn time, τ_{MR}, which is considered by the market to be acceptable to the customers. This market reference can be interpreted as a benchmark the managers use to evaluate the competitiveness of their firms. This benchmark is assumed to be exogenous and fixed. Given $\hat{\lambda}_{jt}$ and τ_{MR}, managers can determine their desired service capacity, $\dot{\mu}_{jt}$, by using Equation 2.8. Rewriting this equation in terms of these three variables we have:

$$\tau_{MR} = \frac{\hat{\lambda}_{jt}}{\dot{\mu}_{jt}^{2}} + \frac{1}{\dot{\mu}_{jt}} \tag{3.3}$$

Note that this is a second order equation for which there are two possible solutions. Nevertheless, due to the nature of the problem it is impossible to have negative arrival and service rates. Hence, we only consider the positive solution of Equation 3.3. This solution is given by:

$$\dot{\mu}_{jt} = \begin{cases} 0 & if\ \hat{\lambda}_{jt} = 0 \\ \dfrac{1+\sqrt{1+4*\tau_{MR}*\hat{\lambda}_{jt}}}{2*\tau_{MR}} & if\ \hat{\lambda}_{jt} > 0 \end{cases} \tag{3.4}$$

We assume $\dot{\mu}_{jt}$ to be the service capacity which managers consider as sufficient to satisfy the customers' needs regarding expected sojourn times. Then, the aim of managers is to adapt their available service capacity (μ_{jt}) to their desired service capacity. They must therefore decide when and how much capacity to add or remove. Nonetheless, once the adjustment decision has been made, its implementation process does not materialise immediately. In fact, when managers decide how much capacity they wish either to add or remove, there is usually a lag between the moment they take their decision and it is implemented. van Ackere et al. (2010) model this type of service systems, where the capacity adjustment involves either an implementation or dismantling time, by applying system dynamics (SD). Examples of this kind of delays include the delivery delay entailed when purchasing new machines; the time required to build new infrastructure; the period for training new employees; and the legal notice period to lay off staff.

According to Figure 3.1, the capacity adjustment entails either to increase (c.f. capacity orders) or decrease (c.f. capacity retirements) capacity. When managers decide how much capacity to add (x_t), these orders accumulate as capacity on order (μ_{jt}^{+}) until they are available for delivery. Once the delivery time (d^{+}) expires the ordered capacity is available for service. That is, order x_t is fulfilled in period $t+d^{+}$. Assuming that there is no capacity on order at the beginning of the simulation, i.e. $\mu_{jt}^{+} = 0$ for $t \leq 1$ and $x_t = 0$ for $t < 1$, the cumulative orders are given by:

$$\mu_{jt}^{+} = \sum_{k=t-d^{+}+1}^{t-1} x_k \quad t > 1 \tag{3.5}$$

Similarly, when the capacity adjustment implies removing capacity, the capacity managers decide to withdraw (y_t) is designated as capacity to be retired (μ_{jt}^{-}). This capacity remains available for customers until the dismantling time (d^{-}) expires, i.e. y_t is removed from the service capacity in period $t + d^{-}$. Assuming that there are no previous retirement decisions at the start of the simulation, i.e. $y_t = 0$ for $t < 1$ and hence $\mu_{jt}^{-} = 0$ for $t \leq 1$, the capacity retirements accumulate as follows for $t > 1$:

$$\mu_{jt}^{-} = \sum_{k=t-d^{-}+1}^{t-1} y_k \tag{3.6}$$

The capacity decisions that have not yet been implemented ($\mu_{jt}^{\pm}$) are given by:

$$\mu_{jt}^{\pm} = (\mu_{jt}^{+} - \mu_{jt}^{-}) \tag{3.7}$$

Regarding the way managers decide the capacity adjustment they require at their facilities (x_t and y_t), we propose a heuristic which enables managers to know how much capacity either to add or remove and when to do so. This heuristic considers that the learning process explained above enables managers to estimate their required capacity adjustment, which managers can decide to carry out as fast or slow as they wish.

The required capacity adjustment depends on the gap managers observe between their desired capacity ($\dot{\mu}_{jt}$) and the service capacity, which they perceive to have currently. When this gap is positive, new capacity orders will be placed, whereas new capacity retirements will be carried out when the gap is negative.

The current available service capacity (μ_{jt}) and the managers' previous decisions, which are still in the process of implementation ($\mu_{jt}^{\pm}$), make up the capacity that managers are expecting to have in service if no further changes are decided. Nevertheless, managers do not necessarily keep in mind all their previous decisions, which have not been yet implemented. Denoting by ψ the proportion of the not yet implemented capacity adjustment, which managers remember, we obtain that the service capacity they perceive to have at time t, is given by:

$$\bar{\mu}_{jt} = \mu_{jt} + \psi * \mu_{jt}^{\pm} \tag{3.8}$$

where ψ is nonnegative and less than or equal to 1. We call ψ the "coherence factor" of managers. If managers are rational when making capacity decisions they should take into account their previous decisions, which are still in process of execution. In this case, $\psi = 1$. Otherwise, if they only account for part of these past decisions, $\psi < 1$.

By computing the difference between the service capacity, which managers consider they currently have, and their desired capacity, we obtain the required capacity adjustment ($\Delta\mu_{jt}$):

$$\Delta\mu_{jt} = \dot{\mu}_{jt} - \overline{\mu}_{jt} \tag{3.9}$$

Managers may make this adjustment as fast or as slow as they wish. That is, we assume that managers can be prudent when taking their decisions. The second element of the heuristic tackles this issue. Let ζ be the speed at which managers decide to adjust capacity, i.e. how fast they decide to either add or remove capacity. Then, when the desired capacity exceeds the current capacity, which managers perceive to have, they decide to add capacity and the ordered capacity (x_{jt}) will be:

$$x_{jt} = \zeta * (\dot{\mu}_{jt} - \overline{\mu}_{jt}) \qquad \textit{if } \dot{\mu}_{jt} > \overline{\mu}_{jt} \qquad \text{with} \quad 0 \leq \zeta \leq 1 \tag{3.10}$$

while if this decision implies to withdraw capacity, the capacity to be retired is:

$$y_{jt} = \zeta * (\overline{\mu}_{jt} - \dot{\mu}_{jt}) \qquad \textit{if } \dot{\mu}_{jt} < \overline{\mu}_{jt} \qquad \text{with} \quad 0 \leq \zeta \leq 1 \tag{3.11}$$

ζ is small when managers take their decisions slowly, i.e. they are prudent decision makers. On the contrary, a high ζ implies that managers take actions quickly, i.e. they are aggressive decision makers.

To summarise the managers' dynamics: The more customers patronise a facility, the higher its manager's perception of the arrival rate is. High (low) managers' expectations increase (decrease) their desired service capacity. The higher (lower) the desired service capacity the more capacity the managers order (remove). With a delay, the capacity orders will increase the service capacity, while the capacity retirements will decrease it. This will affect the number of customer arriving at that facility. These dynamics yield the two reinforcing loops illustrated in Figure 3.1.

The feedback between customers and managers' actions generate different collective behaviours. Indeed, any action of managers on their service capacity will cause changes in the number of customers patronising their facility and vice versa. Moreover, the multiple delays involved in the system influence the agents' decisions and the resulting behaviours. Our aim in this chapter is precisely to study the collective behaviours which emerge from the interaction between customers and managers and the influence of delays on their decisions.

3.3 Customers' and Managers' Profile

Customers and managers can be characterised depending on their decision parameters. Concerning customers, their profile is described using their coefficients of expectations. In the previous chapter we described customers' profile in accordance with their risk-aversion level and the importance they give to new information to update their memory. Now we only consider risk-neutral customers, but we assume that customers can give a different weight to their own information and the one provided by their best performing neighbour. Considering their attitude towards new information, customers can be defined as conservative, hesitant or reactive. We say that customers are conservative or reluctant regarding new information when they have more confidence in their memory than in recent experiences (i.e. α or β high). When the contrary occurs, we call customers reactive (i.e. α or β low). When a roughly equal weight is given to memory and to the new information, we call hesitant customers (i.e. α or β intermediate). Table 3.1 summarises the type of customers according to the weight they give to their memory and depending on the source of information (i.e. own-information or that of their neighbours).

Definition of customer types	**Parameters**	
	w.r.t. their own experience	**w.r.t. the information provided by neighbours**
Conservative customers	$\alpha \geq 0.7$	$\beta \geq 0.7$
Hesitant customers	$0.3 < \alpha < 0.7$	$0.3 < \beta < 0.7$
Reactive customers	$\alpha \leq 0.3$	$\beta \leq 0.3$

Table 3.1 Customer types according to their coefficient of expectations

Like the customers, managers can also be characterised according to their attitude towards new information when updating their perceptions. But, additionally, we can describe managers according to their coherence when taking decisions and the speed at which they implement these decisions. Table 3.2 gives the different characterisations of managers depending on their attitudes when updating perceptions and making decisions.

Definition of manager types	**Parameters**
Conservative managers (Slow decision maker):	
- w.r.t. the speed at which they update their perceptions	$\delta \geq 0.7$
- w.r.t. the speed at which they take their decisions	$\zeta \leq 0.3$
Hesitant managers (Moderate decision maker):	
- w.r.t. the speed at which they update their perceptions	$0.3 < \delta < 0.7$
- w.r.t. the speed at which they take their decisions	$0.3 < \zeta < 0.7$
Reactive managers (fast decision maker):	
- w.r.t. the speed at which they update their perceptions	$\delta \leq 0.3$
- w.r.t. the speed at which they take their decisions	$\zeta \geq 0.7$
Regarding the coherence factor managers can be:	
- fully rational	$\psi = 1.0$
- almost rational	$0.7 < \psi < 1.0$
- slightly irrational	$0.4 \leq \psi \leq 0.7$
- moderately irrational	$0 < \psi < 0.4$
- completely irrational	$\psi = 0$

Table 3.2 Manager types according to the parameters which describe their capabilities.

The agents' profiles described above will be used throughout the remaining of the chapter to explain the behaviour of the system. Likewise, we will refer to the aforementioned attributes when discussing the sensitivity analysis regarding the decision parameters allocated to customers and managers.

3.4 Simulation Results

This section describes the simulation analysis and several experiments performed using the model outlined above. We first present an analysis of four examples of typical behaviours exhibited by the system. Then, we carry out some experiments to study the sensitivity of the behaviour with regard to the decision parameters involved into the model. Specifically, we evaluate how the system responds to changes in the parameters which determine the agents' decision rules. Additionally we examine how the sojourn time benchmark impacts the average sojourn time of the system.

Due to the number of parameter the model has, we limit our simulation analysis to evaluating the dynamics of a system which is configured with 3 facilities and 120 customers (i.e. a one dimensional lattice of 120 cells, where each cell can take exactly one of three states). Each facility is initially provided with a service capacity of 5 customers per time unit, a manager and its own queue. Agents are endowed with an initial memory (i.e. expected average sojourn times for customers and expected arrival rates for managers). We interpret this initial memory as the knowledge customers and managers have about the historical events of the system. This initial memory is randomly allocated to the agents using a uniform distribution, whose maximum and minimum values are respectively 10% above and below the sojourn time of the Nash equilibrium. Given that all facilities have the same service capacity at the beginning, the Nash equilibrium occurs when customers are split equally among the three facilities, i.e. 40 customers patronising each. This distribution yields an average sojourn time of 1.8 time units.

We assume the implementation and dismantling delays involved in the managers' decisions to be fixed and equal to 4 and 2 periods, respectively. That is, once managers decide to increase capacity, this order will be delivered 4 periods later. Similarly, when they decide to reduce capacity, the capacity to be retired will still be available for service during the next 2 periods.

We develop and simulate the model using the numerical computing environment MATLAB 7.9 (The MathWorks TM 2009). We also use MATLAB's statistical toolbox to compute statistics about the performance of the agents and the system (e.g. average sojourn time at each facility and for the system). We use Stata (StataCorp.2011, 2011) to test the statistical hypotheses related to the performance of the facilities.

Owing to limitations of computational capability and the difficulty of visualisation of results in more than 8 dimensions (e.g. $\delta 1$, $\delta 2$, $\delta 3$, α, β, ψ, ζ, τ_{MR}, W are some examples of parameters and variables which determine the dimension of the model), we assume that the three managers are configured with different coefficients of expectations (i.e. $\delta 1$, $\delta 2$, $\delta 3$), but with the same coherence factor (ψ) and decision making speed (ζ). The other dimensions will be studied in further work. Although, MATLAB always performs the operations and keeps the numbers in a precision of 16 decimal digits (The MathWorks TM 2009), we have performed the analysis based only on the first 4 decimal digits.

3.4.1 Typical Behaviours

Here we present the analysis of four examples of typical behaviours generated by the system. These behaviours were obtained running the model with the same initial memory allocated to the agents (i.e. using the same random seed), but varying some of the managers' behavioural parameters. Specifically, we change the coherence factor (ψ) and the decision making speed (ζ).

Agents were initialised with the parameters shown in Table 3.3. Each simulation run consists of 200 periods. The results are studied at the micro and macro-levels. By micro-level, we refer to the individual level, that is, we analyse the way customers' and managers' decisions evolve. The macro-level refers to an aggregated view of the system in which we analyse its overall performance given the agents' decisions.

Figures 3.2, 3.3 and 3.4 illustrate the different results obtained for the four examples. Figure 3.2 illustrates the evolution of customers' decisions and the service capacity (Micro-dynamics). Figure 3.3 shows the evolution of the managers' decisions and the service capacity during the transient period. Figure 3.4 shows the average, minimum and maximum sojourn times that customers experience at the facilities each period (an aggregated view of the system).

Figures 3.2 (a) (c), (e) and (g) show the space-temporal behaviour patterns generated by the customers' choice of service facility. The customers' decisions during the first four periods are the same for the four examples. They move between the different facilities to experiment and update their expectations. This occurs because the customers and managers' initial memories are the same for these four examples. Additionally, the three facilities are initially setup with the same service capacity. Thus, customers will behave similarly in these examples until a change occurs with regard to the service capacity.

Figures 3.2 (b) (d), (f) and (h) show the evolution of the service capacity at each facility resulting from the managers' decisions. Although managers of identical facilities have the same initial desired service capacity in the four examples (e.g. the manager of facility 2 is allocated with the same initial memory in the four examples), their first capacity decisions vary between these examples. Indeed, the magnitude of these orders also depends on the speed at which managers take their decisions (ζ) (their coherence factor (ψ) plays no role initially since it is assumed that there is no capacity on order in period 1). This speed is different in the four cases. In this sense, managers in example 4 (Figure 3.2 (h)) make the largest capacity order in period 1 since they take decisions faster (ζ = 0.8). The magnitude of the following capacity adjustments will depend on the system dynamics, i.e. the interactions between customers and managers.

Parameter	**Examples**			
	(A)	**(B)**	**(C)**	**(D)**
α	0.3 (Reactive)	0.3 (Reactive)	0.3 (Reactive)	0.3 (Reactive)
β	0.7 (Conservative)	0.7 (Conservative)	0.7 (Conservative)	0.7 (Conservative)
$\delta 1$	0.8 (Conservative)	0.8 (Conservative)	0.8 (Conservative)	0.8 (Conservative)
$\delta 2$	0.5 (Hesitant)	0.5 (Hesitant)	0.5 (Hesitant)	0.5 (Hesitant)
$\delta 3$	0.2 (Reactive)	0.2 (Reactive)	0.2 (Reactive)	0.2 (Reactive)
ψ	0.5 (Slightly irrational)	0.8 (almost rational)	0.33 (moderately Irrational)	0.7 (Slightly Irrational)
ζ	0.2 (Slow)	0.5 (Moderate)	0.66 (Moderate)	0.8 (Fast)
Random Seed	12225	12225	12225	12225

Table 3.3 Parameter values used for the simulation runs.

3.4.1.1 Individual behaviour

Changes in service capacity depend on both the managers' decisions and the implementation delays of these decisions. In period 1, managers use their initial memory to compute their initial desired service capacity. This estimate is lower than their current service capacity since it is based on the benchmark of sojourn time (2 time units), which is higher than the average sojourn time corresponding to their expected arrival rate given the initial conditions (10% above the Nash equilibrium of 1.8 time units). Consequently the managers' first period decision implies a slight reduction in service capacity (Figure 3.3) which, given the dismantling delay, materialises in period 3 as show in Figure 3.2.

After period 4, the behavioural patterns exhibited in Figures 3.2 can be explained in two different phases. The first phase exhibits a transient period whereby customers and managers try to learn from the system. In the previous chapter we explained that the length of this period depends on the random initial conditions. During this period, customers move through all facilities attempting to find the facility with the lowest average sojourn time; while managers attempt to adjust their service capacity to satisfy the customers. After the transient period, a collective behaviour of the agents emerges and the system exhibits a certain stability. When this occurs, we say that the system has reached a steady-state.

The customer behaviour during the transient period has been extensively discussed in the previous chapter and in Delgado et al (2011a). The managers react to customer behaviour by adjusting their service capacity. In addition to the system dynamics (i.e. the interaction between customers and managers), the kind of adjustment (i.e. increasing or decreasing capacity) managers make and its magnitude depend on the managers' profile. Both elements depend on the coefficient of expectations (since it determines the desired service capacity) and the consistency factor (it influences the capacity, which the manager perceives to have). However only the decision making speed affects the magnitude of the adjustment (see Equations 3.10 and 3.11). For instance, we observe in Figure 3.3 that for all the examples, managers 1 (green line) and 2 (blue line) are more moderate than manager 3 (red line) when adjusting service capacity during the first eight time periods. Specifically, the service capacity of manager 3 is significantly reduced compared to the one of the other two managers. Unlike manager 3, managers 1 and 2 are more conservative regarding the new information (see δ coefficients). Likewise, Figure 3.2 (c.f. the evolution of customers' choice) shows that the number of customers patronising facilities 1 (grey cells) and 2 (black cells) during the first six time periods are somewhat more stable than for facility 3 (white cells). Hence, the adjustments made by managers 1 and 2 are gentler than those made by their counterpart at facility 3.

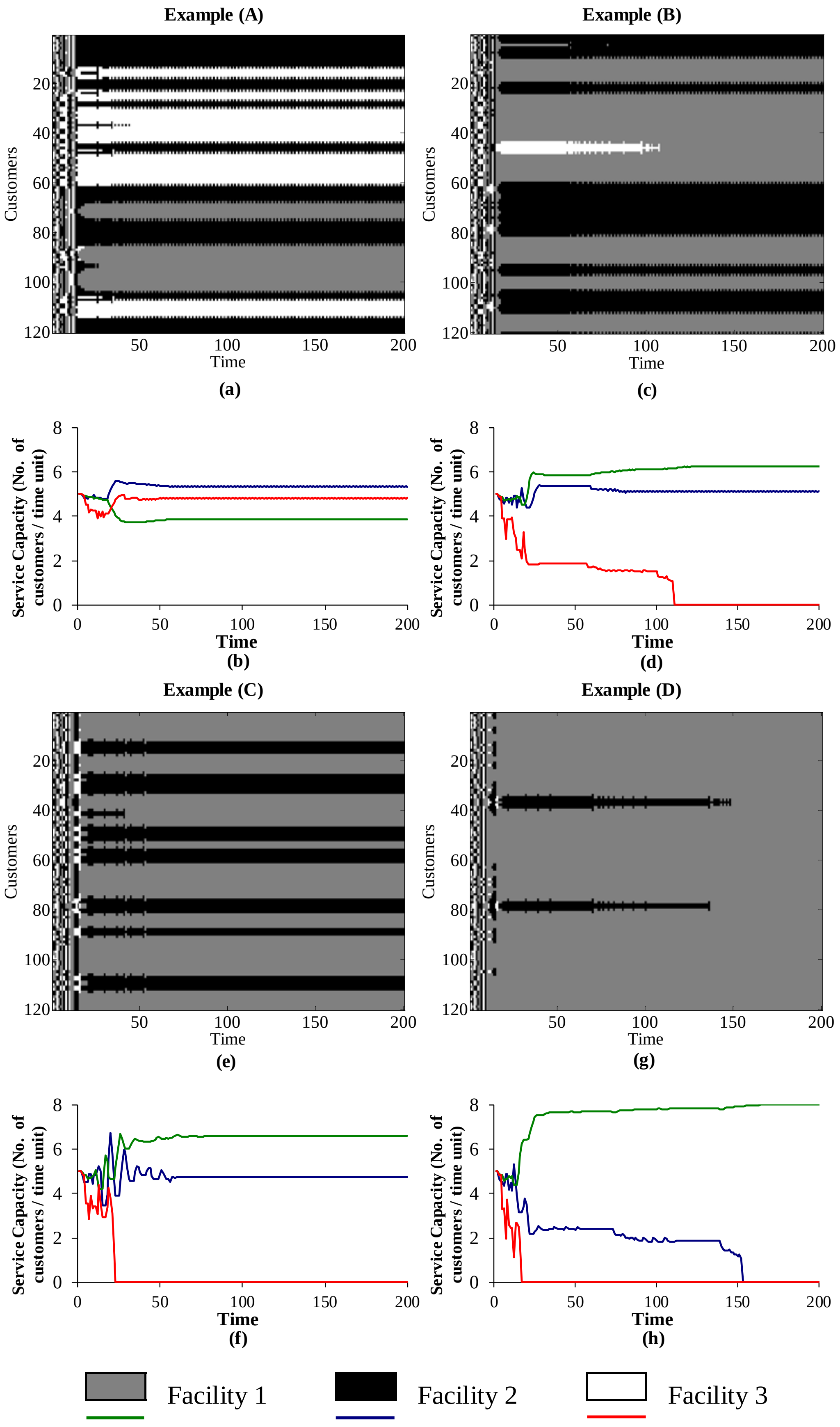


Figure 3.2 Space-time evolution patterns of customers' choice of service facility and the service capacity evolution of each facility depending on the parameters involved in the managers' decision rule (see Table 3.3).

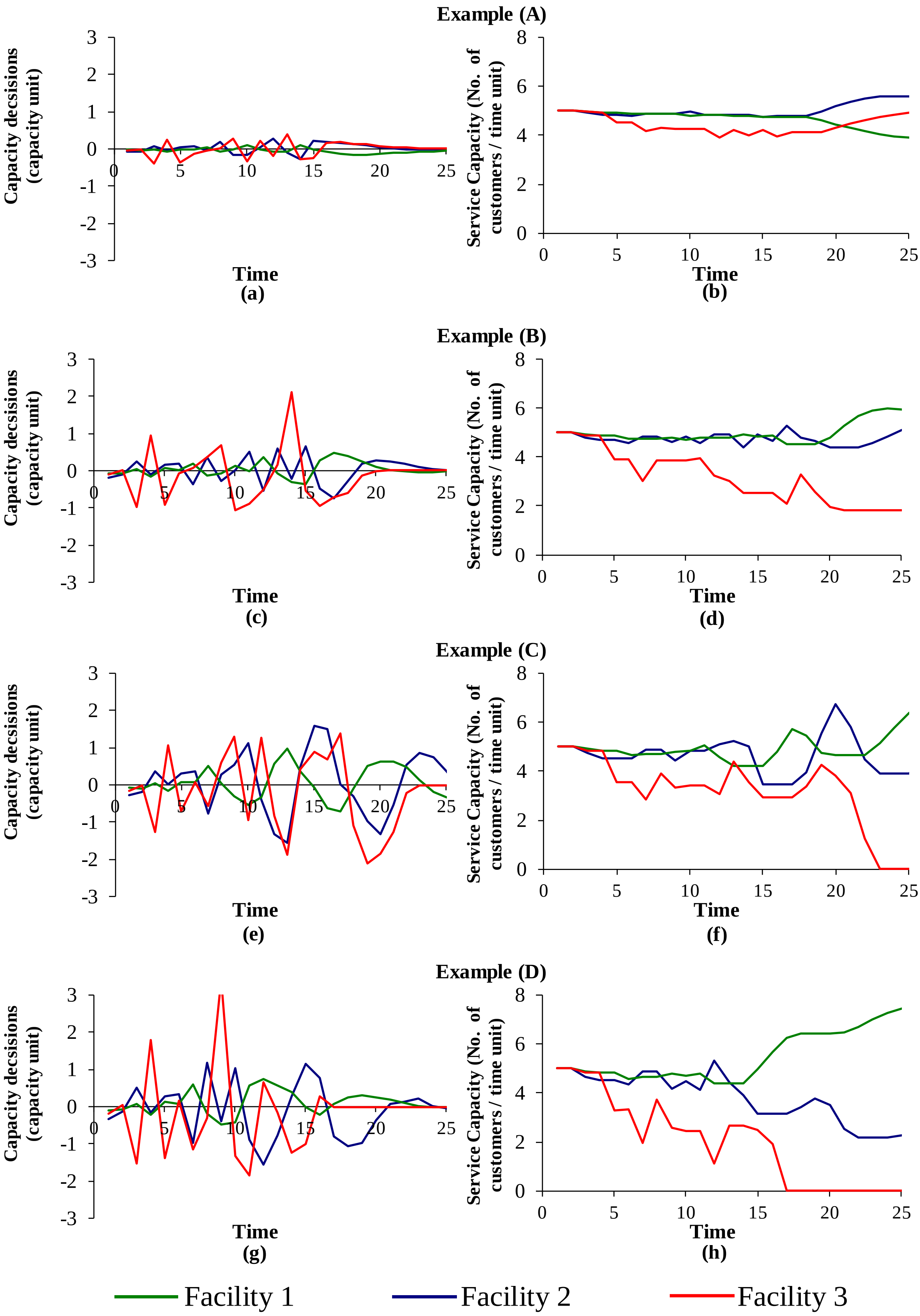


Figure 3.3 Managers' capacity decisions and the service capacity evolution during the first 25 time periods of the transient period for the four examples specified in Table 3.3.

The behaviour of manager 3 at the beginning can be explained by two factors: the customers' experience at facility 3 during the first three time periods and the manager's attitude toward new information. Given the initial conditions he starts by deciding to remove some capacity (See the red line in figures in the left column of Figure 3.3). In periods 1 and 3, facility 3 is the most patronised (see white colour in Figures 3.2 (a) (c), (e) and (g)), while it is the least patronised in period 2. As the arrival rate at facility 3 is very low in period 2, his manager, who is the most reactive regarding the new information when updating his memory ($\delta 3$ = 0.2), decides to withdraw even more capacity in period 3, as shown in the left column of Figure 3.3 (red line). At that time customers come back to this facility. Consequently in period 4, manager 3 takes the decision to increase capacity, while customers ignore his facility. So, in period 5 when the capacity removing decision taken in period 3 materialises (see right column of Figure 3.3), he overreacts by removing even more capacity. It is worth recalling that the decision making speed differs across the four examples and thus the magnitude of the managers decisions differs from one case to another. Given the delivery and dismantling delays (4 and 2 time periods, respectively), the decision taken in period 5 to decrease capacity materialises before the decision to increase capacity taken in period 4, as can be seen in Figure 3.3 (b), (d), (f), and (h). Additionally, most of the capacity orders placed by manager 3 are partially offset by his subsequent retirement decisions. After period 8, when the capacity order placed by manager 3 in period 4 is the first to materialise (see Figure 3.3 (b), (d), (f), and (h)), the behaviour of this manager changes notably in the four examples: he makes more aggressive decisions. Below we discuss in more detail the behaviour exhibited by both managers and customers in each example.

Figure 3.2 (a) and (b) show an example where the three facilities remain open during the whole simulation period. In this case, the managers only account for 50% of their not-yet implemented decisions when taking the next one. Similarly, they are very prudent and therefore they make decisions slowly (at a rate of 20% of the desired adjustment). The transient period for this example lasts about 45 periods. From then on, customers exhibit a quasi-stable behaviour in which groups of customers loyal to a facility emerge and the customers located on the borders of these groups keep alternating between two facilities (the facilities patronised by their neighbours), as shown in Figure 3.2 (a). Nevertheless, the service capacity takes a bit more time to stabilise, but this is normal given the delays involved in the decision process. Once the system stabilises, facility 2 (black cells in Figure 3.2 (a)) is the most patronised and facility 1 (grey cells) the least. Similarly, these are the facilities with the highest and lowest service capacity, respectively. All the three facilities have almost the same sojourn time

It is worth mentioning that in this example, the service capacity of the three facilities presents the smallest fluctuations through the transient period compared to the other three examples (See Figures 3.2 (b) and 3.3 (b)). Note that the manager of facility 1 keeps a service capacity similar to that of facility 2 and larger than that of facility 3 (red line and white cells) even though this facility has been already totally ignored three times before period 14. Moreover, from that moment on this facility is the least patronised until period 19. In period 13, 112 customers patronise facility 1, which is the maximum for any facility throughout the simulation (Note that at no time do all the 120 customers use the same facility). This is very positive for manager 1, but a very bad experience for his customers. So customers ignore facility 1 in period 14 (see Figure 3.2 (a)), while its manager, given his profile (very conservative regarding new information, moderately irrational and slow decision maker), takes a very small decision to increase the service capacity, as can be seen in Figure 3.3(a). The next time periods, customers are slowly coming back to facility 1 and its manager reacts to this behaviour by gradually reducing the service capacity. This manager keeps the policy of decreasing capacity for the next 20 time periods and his service capacity falls below that of facility 3, as shown in Figure 3.2 (b).

Most of the capacity orders placed by manager 2 are partially offset by his subsequent retirement decisions between periods 8 and 18. Unlike manager 1, he decides to slightly increase capacity between period 15 and 21. These decisions materialise from period 19 onwards and enable him to have the largest facility.

In period 15, manager 3 starts to adding small amounts of capacity each period (red line) and his facility starts to inherit some customers who have experienced high sojourn times at facilities 1 and 2 in the previous two time periods. From then onwards, facility 3 is the second most patronised. This encourages manager 3 to order more service capacity during the next ten periods. These orders start being fulfilled from period 20 onwards, as shown in Figure 3.2 (b) and Figure 3.3(a) and (b). From then onwards, this facility has the second highest service capacity.

The other three examples illustrate cases where at least one facility closes. This occurs logically when customers ignore a facility for a long time and its manager consequently reduces capacity, ultimately closing down the facility. Nevertheless, irrational managers do take decisions which lead to a shutdown (capacity equal to zero) even in the presence of customers. The occurrence of either of these two events yields a kind of path dependence (Arthur 1989, Liebowitz and Margolis 1995), in which the system becomes "locked-in" into a monopoly or duopoly situation, as once a facility has closed down, it cannot re-open. The emergence of this phenomenon depends on both the customers' and the managers' previous decisions as well as on the initially allocated memories. A more detailed analysis of how the customers' and the managers' previous decisions can lead to such path dependence is given below.

Figures 3.2 (c), (d), (e) and (f) illustrate two examples in which one of the three facilities closes down. In both cases, facility 3 shuts down because the customers walked out.

In example (B) (Figures 3.2 (c), (d)), managers are almost rational ($\psi = 0.8$) and decide moderately fast ($\zeta = 0.5$). The transient period of this case can be explained in two phases. The first phase is similar to the previous example: customers move through all facilities during a certain time and managers overreact to this behaviour by increasing and decreasing service capacity (See Figure 3.3 (c)). Given that managers decide more quickly than in example (A), the facilities' service capacity show greater fluctuations. This phase takes around seventeen time periods. During this period, the manager of facility 3 sharply reduces his service capacity (red line). This strongly affects the experience of customers at this facility and therefore their memory. In period 13, most customers decide to join facility 3 after having had a bad experience at the other two facilities during the previous two periods. Given that facility 3 has much less service capacity than the other two, the experience at this facility is worse and the impact on the customers' memory is much stronger. So, this facility is ignored in period 14 and only three customers come back the next period. Even though manager 3 makes a large capacity order in period 14 (based on the demand of period 13), only two more customers join this facility in period 18 when such an order materialises. From that moment on the system exhibits an interesting behaviour in which the customers' and managers' decisions seem to have reached the Nash equilibrium, since customers do not switch facility and the service capacities seem to remain constant. However, the service capacity of the facilities does change, but the changes are very small and graphically unobservable. Additionally, the system can only reach the Nash equilibrium when the average sojourn time of all facilities is exactly equal the benchmark (2 time units) and this is not the case. Thus, these small changes in service capacity affect the average sojourn time and, in turn, the customers' memory. Two customers of facility 3 are the first to break this false equilibrium in period 55. From then onwards, they irregularly alternate between this facility and facility 1 until they persuade their neighbours to leave facility 3 forever in period 107. This causes manager 3 to shut down his facility three time periods later. Then, an almost-stable behaviour emerges. Customers and managers behave as in the previous example when the system reached steady-state.

Figures 3.2 (e) and (f) show another example where only two facilities remain open. During the transient period, the service capacity exhibits more and higher fluctuations than in the previous two examples. This is because managers are moderately irrational (ψ = 0.33), which means that they only consider a small portion of their previous decisions, which are not yet implemented, when making future decisions. Moreover, they decide faster than the managers of the previous two examples (ζ = *0.66*). The closing down of a facility occurs much earlier during the transient period. In period 16, facility 3 achieves its lowest level of service capacity so far and it is the most patronised facility, resulting in a bad experience for customers. As a result, customers decide to ignore this facility for the next time periods and its manager shuts it down six time periods later.

The main difference between this example and the previous one lies in the system behaviour during the steady-state period. Indeed, we observe that the system reaches the Nash equilibrium with two facilities. After time 54, seven groups of customers loyal to facility 1 and seven groups to facility 2 emerge. The more customers patronise facility 1 which has the highest service capacity. The managers' decisions take more time to reach equilibrium than customers' choices. This occurs in period 100, when both facilities have enough service capacity to guarantee that their customers will always experience an average sojourn time equal to the benchmark.

In the last case shown in Figures 3.2 (g) and (h), only one facility remains open. This is facility 1 (grey cells and green line), whose manager is the most conservative. Facility 3 (white cells and red line) is the first to close. Its manager significantly reduces the service capacity during the transient period. As a result, this facility is ignored four times during the first ten time periods. Similarly, this facility is twice very crowded during these ten periods. Hence, manager 3 decides to sharply increase his service capacity, but he later counteracts this decision by decreasing capacity (see Figure 3.3 (g)), and given the length of the delays, the capacity retirements materialise before the capacity orders. Then, in periods 13 and 14 facility 3 is ignored again. This motivates its manager to reduce the service capacity over the next time periods. Consequently, this facility is closed in period 17, even though a few customers had come back to this facility in period 15. Facility 2 is the other one which closes. The behaviour of this manager is similar to the one exhibited by manager 3 in example 2. He closes his facility because customers gradually leave.

3.4.1.2 Aggregated behaviour

We use the average sojourn time customers experience at the facilities as a measure which enables us to analyse the global system performance. Figure 3.4 shows the evolution of the weighted average sojourn time of the system and the minimum and maximum sojourn time experienced by customers at each facility for the same four examples discussed above. The weighted average sojourn time is computed using Equation 2.10 (see Chapter 2). Additionally the weighted average sojourn time in steady state is computed for the last 40 periods.

We can observe in Figure 3.4 that in all the cases, the average sojourn time in steady-state converges to the benchmark (2 time units). This behaviour is in keeping with the goal of the managers of satisfying the customers' needs, which are represented by the reference average sojourn time for the market, i.e. the benchmark (τ_{MR}). Facilities which do not approach this benchmark soon get out of business, as was illustrated in Figure 3.2. This aspect enables us to conclude that the benchmark of the average sojourn time acts as an attractor point in the system for the agents' decisions.

The average sojourn time of examples (a) and (b) seems to be constant. However, this measure actually presents a sustained oscillation during the steady-state period, as can be seen from the minimum and maximum values. Note that the oscillations in the second example have greater amplitude than in the first one. Nevertheless, given the almost-stable behaviour of these decisions, the average sojourn time of customers at the facilities fluctuates over time with a very small constant amplitude as shown in Figure 3.4.

In the context of queueing systems, the Nash equilibrium implies that customers do not wish to change facility. This occurs only when the average sojourn time is the same at all facilities. Figure 3.2 (e) shows that in the third example customers remain at the same facility in steady-state and Figure 3.4 (c) shows that the maximum and minimum average sojourn times equal the weighted average sojourn time of the system (i.e. all customers experience the same average sojourn time in steady state). Hence, we can conclude that in the third case the system has really reached the Nash Equilibrium.

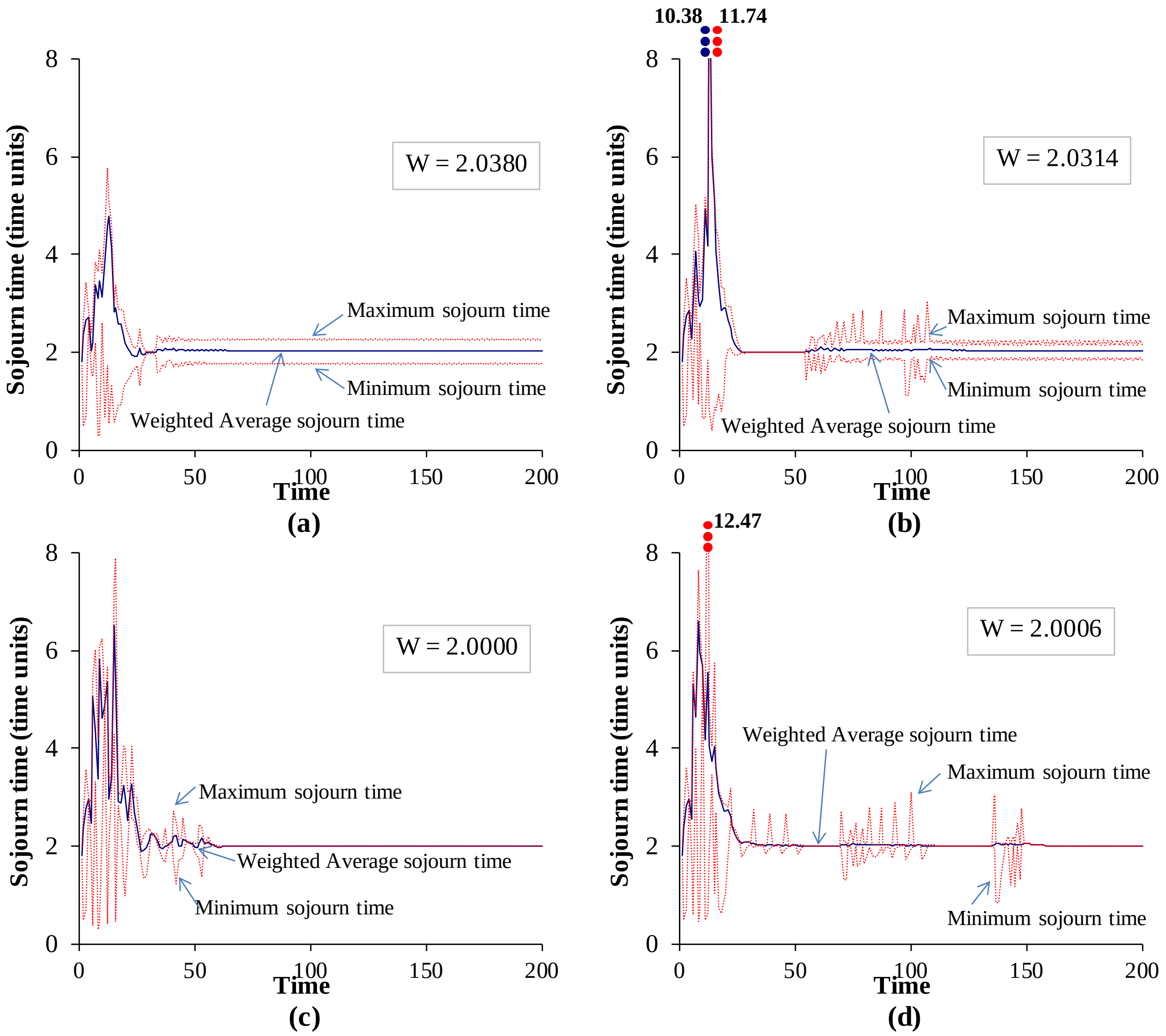


Figure 3.4 Average sojourn time evolution in four different examples depending on the parameters of the managers' decision rule.

In the last example only one facility remains open. The average sojourn time in steady state is still above the benchmark (2 time units) at time 200. This is due to the long transient period exhibited by this system. When the transient period ends, the manager takes much more time to stabilise his decisions. Nevertheless, the average sojourn time of this facility will also converge to the benchmark but after a much longer time.

The four examples described above are the most typical behavioural patterns we can observe when simulating the model for different combinations of the model parameters (i.e. managers, customers and system parameters) and using different random seeds to allocate the initial memory to the agents. As general conclusion of the behaviour observed in these four typical cases, we can say that conservative managers usually achieve a greater market share.

Other behaviours, though rare, can occur:

1) The most conservative manager closes his facility (e.g. manager 1 in the previous cases). In the next section we show evidence regarding this case.
2) The facility of the most reactive manager is the only one remaining open.
3) A Nash equilibrium with the three facilities remaining open. This tends to occurs when the system is set up with extreme values of the coefficients of expectation (e.g. $\alpha = \beta = 0.9; \delta 1 = \delta 2 = \delta 3 = 0.1$)
4) The transient period lasts for a very long time, that is, the system takes more than 500 time periods to reach a steady-state.
5) All facilities close down. This usually occurs in extreme cases when managers are close to fully irrational (e.g. $\psi < 0.1$) and very slow in taking decisions ($\zeta > 0.8$).

Next we perform a sensitivity analysis of the impact of some model parameters on the managers' and customers' decisions.

3.4.2 Simulation Experiments

This section describes the analysis of simulation experiments performed using the model outlined above. These experiments are discussed in terms of the possible scenarios which we can obtain when simulating the model as explained in Table 3.4. In all the experiments we run the model for 500 time periods and compute the steady-state value over the last 40 time periods.

We carry out 3 experiments. The first two concern the managers' parameters (δ, ψ and ζ). The third assesses the sensitivity of the average sojourn time of the system to the benchmark (i.e. the market reference, τ_{MR}). The customers' parameters (α and β) are assumed to be fixed. Thus, their analysis only applies for systems whose customers are set up with $\alpha = 0.3$ (slightly reactive regarding their own information) and $\beta = 0.7$ (slightly conservative regarding their neighbours' information).

Colour code	Numerical code	Scenario
⬜	0	All facilities close
🟦	1	Facility 1 is the only one open
🟦	2	Facility 2 is the only one open
🟦	3	Facility 3 is the only one open
🟩	12	Facilities 1 and 2 remain open, while facility 3 closes
🟩	13	Facilities 1 and 3 remain open, while facility 3 closes
🟩	23	Facilities 2 and 3 remain open, while facility 3 closes
🟨	123	All facilities remain open

Table 3.4 Possible scenarios generated by simulating the model.

3.4.2.1 Path dependence and the influence of the managers' parameters on the closing of a facility.

The first experiment is aimed at studying the impact the managers' parameters have on the number of facilities remaining open over time. Using the values {0.2, 0.5, 0.8} we consider all combination of the coefficients of expectations of the managers ($\delta 1$, $\delta 2$, $\delta 3$), i.e. twenty-seven combinations altogether. For each combination we simulate the model varying the coherence factor (ψ) and the speed at which managers take their decisions (ζ). These last two parameters vary from zero to one in steps of 0.01. This generates 10,201 combinations of these two parameters for each configuration of the coefficients of expectations, yielding a total of 275,427 runs. Each simulation run is thus a different case which represents a combination of parameters $\delta 1$, $\delta 2$, $\delta 3$, ψ and ζ. Each case is run using the same seed used to obtain the typical behaviours in the previous section, i.e. seed number 12225. This means that the initial conditions allocated to the agents are the same in each simulation.

Figures 3.5 to 3.7 show which facilities are still in operation at time period 500 as a function of the parameters allocated to the managers for these 275,427 cases. Each of the nine panels in Figures 3.5 to 3.7 represents one of the twenty-seven combinations of the coefficients of expectations of managers ($\delta 1$, $\delta 2$, $\delta 3$). For each of these combinations, the model is simulated as a function of the parameters ψ and ζ. In each case, only one scenario of Table 3.4 can occurs, i.e. each case is represented by a cell and each cell takes on a colour from Table 3.4 depending on which scenario occurs. Yellow cells represent the scenarios where the three facilities remain open, while the white ones mean that the three facilities close. The blue shades indicate that only one facility remains open and the green ones that two facilities remain open. When we refer to facility 1 as an open facility, we look for any colour whose numerical code involves “1” (i.e. dark blue, dark green, medium green and yellow) in Table 3.4.

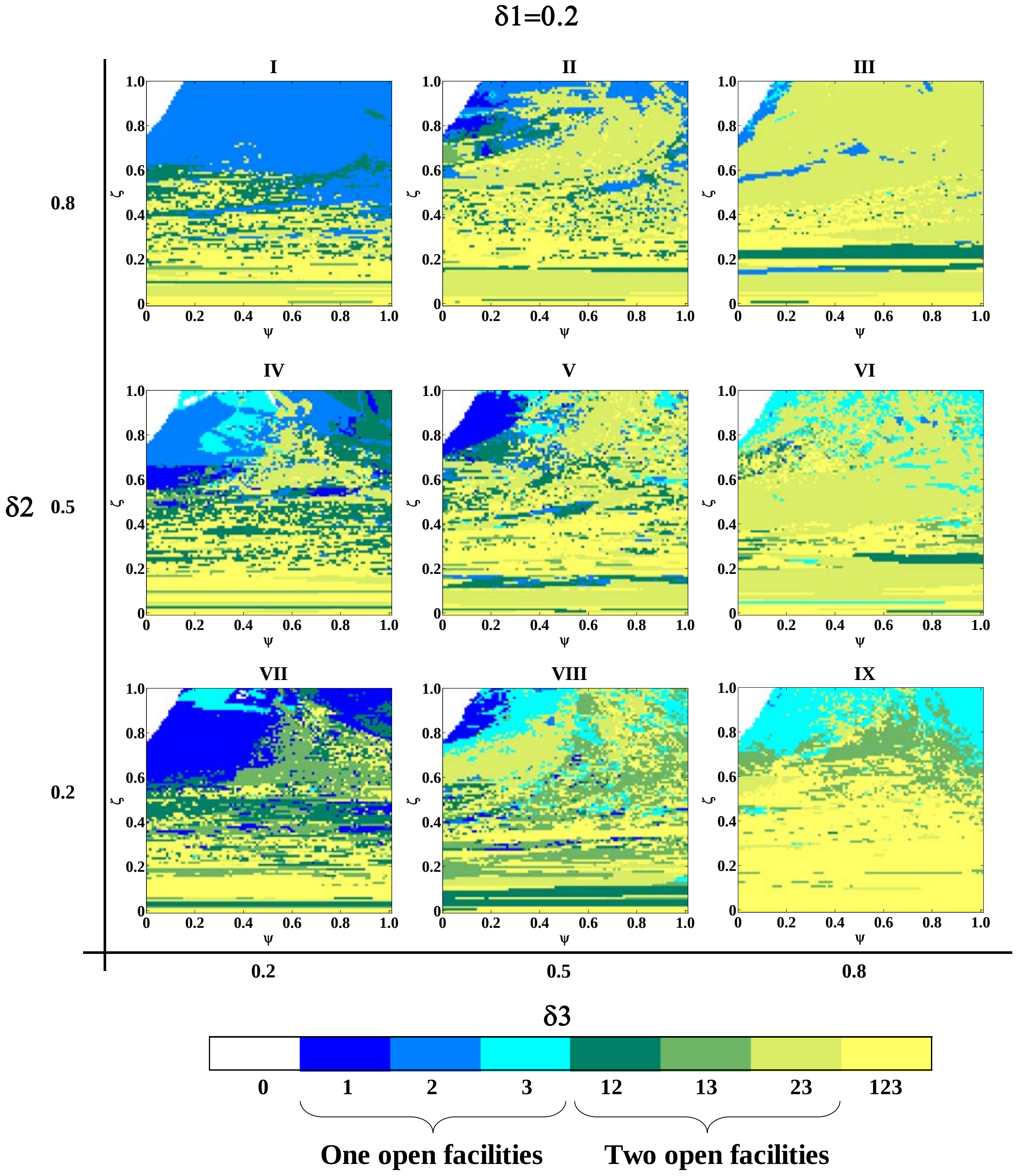


Figure 3.5 Illustrative simulation results of the facilities still open at time 500 for the cases where manager 1 is reactive regarding new information (*δ1* = 0.2), as a function of the coefficients of expectations of managers 2 and 3 (*δ2*, *δ3*), as well as of *ψ* and *ζ*.

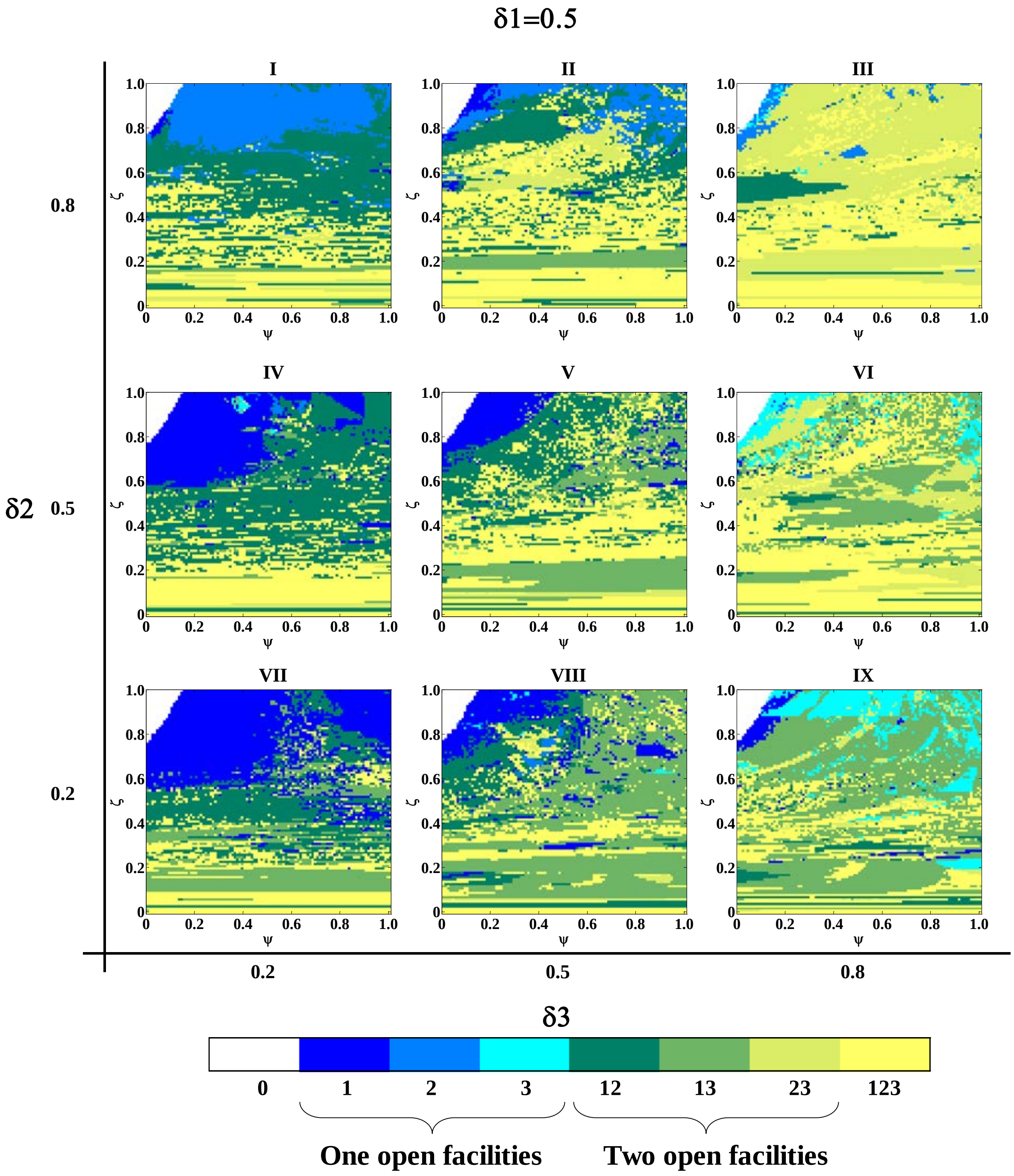


Figure 3.6 Illustrative simulation results of the facilities still open at time 500 for the cases where manager 1 is hesitant regarding new information (*δ1* = 0.5), as a function of the coefficients of expectations of managers 2 and 3 (*δ2*, *δ3*), as well as of *ψ* and *ζ*.

Figure 3.5 illustrates which facilities remain open for the all the cases in which manager 1 is assumed to be reactive regarding new information (*δ1* = 0.2), while the other managers can be either reactive, hesitant or conservatives. Figure 3.6 assumes that manager 1 is hesitant and Figure 3.7 that he is conservative.

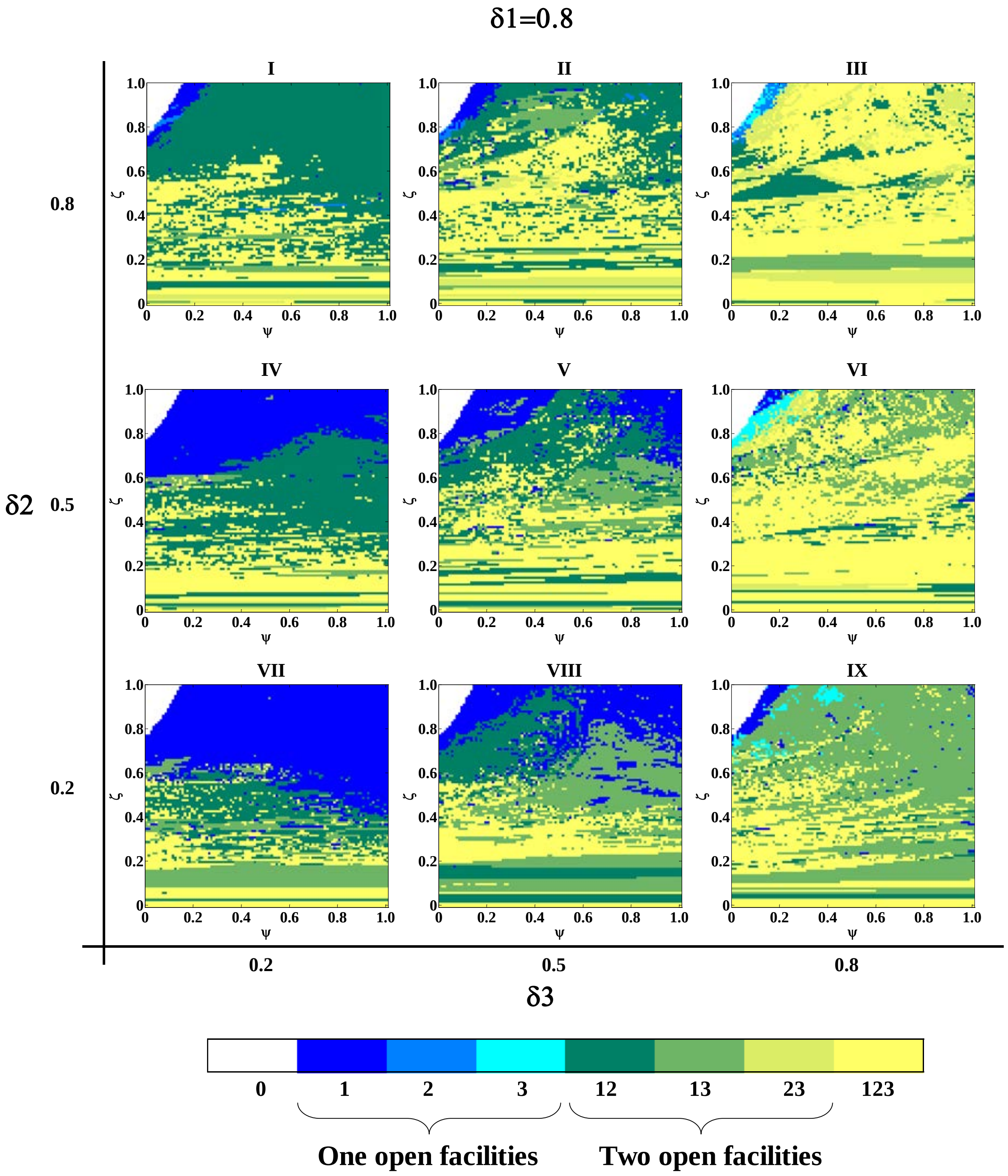


Figure 3.7 Illustrative simulation results of the facilities still open at time 500 for the cases where manager 1 is conservative regarding new information (*δ1* = 0.8), as a function of the coefficients of expectations of managers 2 and 3 (*δ2, δ3*), as well as of *ψ* and *ζ*.

It is worth recalling that these cases are generated using a set of initial memories randomly allocated to the agents (i.e. a random seed). While using a different random seed would yield somewhat different figures, the qualitative discussion provided below remains valid. In other words, while the quantitative results (e.g. the exact percentage of cases that facility 1 remains open) change, the qualitative results (e.g. the most frequent outcome is that facility 1 remains open) remain valid.

The main insight we can draw from Figures 3.5 to 3.7 is that the more conservative a manager (higher values of δ) is compared to his competitors, the larger the percentage of cases in which he keeps his facility open (see panels on the top and on the right of Figure 3.5 and Figure 3.6 as well as all panels in Figure 3.7). For instance, note that in Figure 3.7, where manager 1 is the most conservative, there are very few cells taking the colours medium blue (code 2), light blue (code 3), light green (code 23) or white (code 0).

Most of the monopoly situations (blue shades) occur when a conservative manager, who takes decision fast (ζ >0.6 approximately), faces two reactive managers (see the cases with ζ >0.6 in panels I and IX in Figure 3.5 and panel VII in Figure 3.7). In these cases, the facility of the conservative manager achieves a monopoly. Note that for managers 1 (dark blue in panel VII of Figure 3.7) and 2 (medium blue in panel I of Figure 3.5), there are more cases (i.e. combinations of parameters ψ and ζ) yielding this type of scenario (i.e. monopoly situation) than for manager 3. This illustrates the impact of the choice of random seed and emphasises the phenomenon of path-dependence discussed above. That is, the managers' decisions are strongly "path-dependent", in the sense that the facilities of two or three rival managers, who have the same profile and face the same customers, could evolve completely different each other, depending on the initial conditions.

Panel VII of Figure 3.5, panel V of Figure 3.6 and panel III of Figure 3.7 represent those cases where the three managers have identical profiles. By comparing these three panels we observe that (i) there are more parameter combinations which yield a monopoly position for facility 1 (dark blue) than for the other two facilities; (ii) there are a few parameter combinations in which facility 1 closes down (i.e. there are very few cells taking the colours medium blue, light blue, light green or white); and (iii) the number of cases where the three facilities are still open at time 500 (yellow) increases as managers are more conservative regarding new information.

In some extreme cases, e.g. when managers are close to being fully irrational (e.g. $\psi < 0.1$) and very slow taking decisions (e.g. $\zeta > 0.8$), all the facilities close. This also applies for other set of random initial memories allocated to the agents. However, we do not delve into these cases, since we assume that these extreme conditions are not realistic and would lead to entry by more rational competitors.

In general we can say that the more conservative the managers are, the less influence the coherence factor (ψ) and the decision making speed (ζ) have on the closure of their facilities.

3.4.2.2 Influence of the managers' parameters on the probability of a facility being closed.

The second experiment implies 1,000 iterations (i.e. 1,000 different random seeds) of the model for a number of different combinations of the managers' parameters ($\delta 1$, $\delta 2$, $\delta 3$, ψ, ζ) for the case where $\alpha = 0.3$ and $\beta = 0.7$. In order to validate if 1,000 iterations are enough to draw conclusions about the different scenarios the system exhibits in steady state, we have run 10,000 iterations of the model and extended the simulated time to 10,000 time periods for several parameter combinations. The steady-state period was computed for the last 100 time period. We found that there were no significant differences in the number of facilities closing compared to 1,000 iterations over 500 periods. We therefore assume that 1,000 simulations of the model over 500 time periods is appropriate for our analysis.

Figure 3.8 shows the relative frequency of each scenario described in Table 3.4 for the nine possible combinations of parameters (ψ, ζ) using the three values {0.2, 0.5, 0.8}. The coefficients of expectations of managers and customers are those used in Section 3.4.1 (see Table 3.3: $\delta 1 = 0.8$, $\delta 2 = 0.5$, $\delta 3 = 0.2$, $\alpha = 0.3$ and $\beta = 0.7$). Each combination of parameters (ψ, ζ) is represented by a bar in Figure 3.8. The first three bars (i.e. I, II, and III) illustrate the cases where the managers are slow decision makers ($\zeta = 0.2$) and have different degrees of rationality when accounting for their not-yet implemented decisions (ψ). The next three bars represent those cases where managers are moderate decision makers and the last three those where managers are fast decision makers.

Figure 3.8 indicates that the scenario where the three facilities remaining open (yellow colour) is the most likely when managers take their decisions slowly (see bars I, II and III). This probability decreases as the decision making process is faster and it is close to zero when managers are faster decision makers ($\zeta = 0.8$) and almost rational when accounting for their not-yet implemented decisions ($\psi = 0.8$) (case IX). In contrast, the probability of one facility closing (green shades) is lower when the decision making process is slow (see bars I, II and III) and is the highest when this process is moderate (see bars IV, V and VI). Note in particular that when one facility closes, this is mostly facility 3 (dark green), whose manager is the most reactive (*δ3* = 0.2). Additionally this scenario is the most likely in all the cases in which managers are moderate decision makers (bars IV, V and VI) and in those cases where they are fast decision makers and either slightly irrational or almost rational when considering their not-yet implemented decisions (bars VIII and IX).

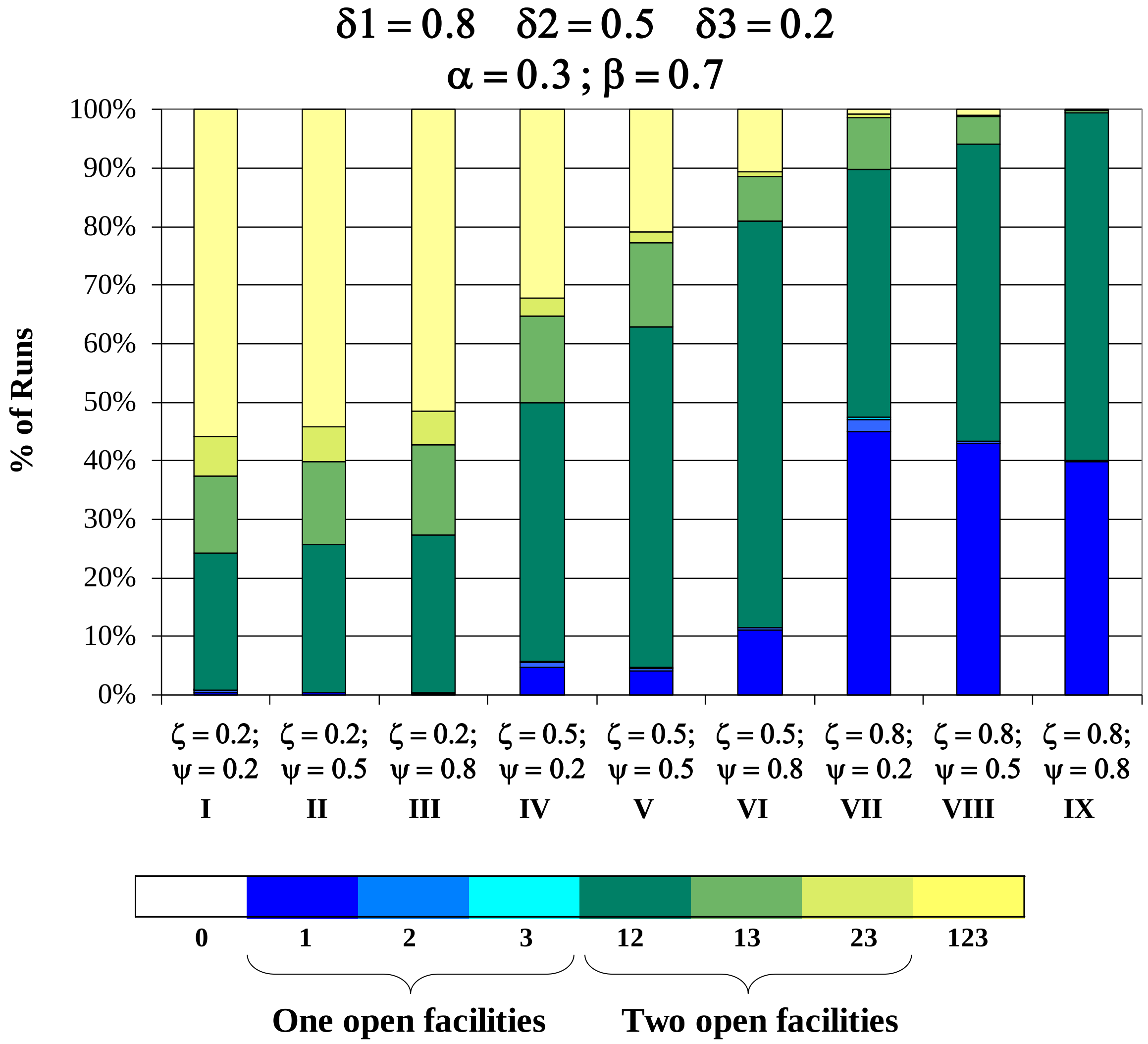


Figure 3.8 Percentage of runs which yield each possible scenario depending on the consistency factor (*ψ*) and the speed factor (*ζ*).

The faster the decision making process, the lower the probability of the most conservative manager (i.e. manager 1, *δ1* = 0.8) being the only one who closes his facility (light green). The probability of the hesitant manager (i.e. manager 2, *δ2* = 0.5) being the only one to close his facility (medium green) is higher when the decision making process is slow (bars I, II and III) and lower when such a process is fast (bars VII, VIII and IX). When the most conservative manager (i.e. manager 1) decides quickly he is more likely to achieve a monopoly position (See medium blue in bars VII, VIII and IX). In the case where managers are almost rational and take decisions fast, the probability of the most conservative manager closing his facility is negligible (less than 0.1%) (bar IX). Finally, note also that the scenario in which the three facilities shut down is very unlikely (less than 0.2%) for this case.

Concerning the market share of the facilities in each scenario, we hypothesise that the most conservative managers' facilities capture a larger market share than the others. We use the Mann-Whitney-Wilcoxon test (MWW) (Newbold 1988), a non-parametric statistical test also called the Mann-Whitney U-test, to assess the null hypothesis that the median of the average arrival rates of two facilities during steady-state are the same. The alternative hypothesis assumes that the median of the average number of customers arriving at the most conservative manager's facility in steady-state is greater than that at the other facility. We have applied a non-parametric test because we do not know the distribution of the data and for some scenarios we have very little data. Parametric tests are not appropriate in these cases. Instead, the MWW-test is appropriate because the distributions have enough symmetry to assume that the median and the average are similar.

Table 3.5 provides the test-statistics of the MWW-test to assess the difference between the medians of the distributions of average arrival rates at each facility in steady-state for the scenarios where at least two facilities remain open, as monopolistic situations are irrelevant in this context. This table contains statistics for the same nine cases of Figure 3.8. For each case and each scenario, this table provides the following information which, as an example, is explained in detail for case I and the scenarios where facilities 1 and 2 remain open (dark green) and where all facilities remain open:

- The relative frequency of iterations in which the average number of customers arriving at the active facilities is the same (referred as "P(equal)" in Table 3.5; for instance, in case 1 when facility 1 and 2 remain open (dark green scenario) 0.85% of the iterations yield that both facilities serve on average the same number of customers in steady state. When all facilities remain open (yellow scenario) not a single iteration yields this event.
- The probability of a given facility (F_x) capturing the largest market share (referred as "P[λ(Fx)=Max(λ)]); for instance, when facilities 1 and 2 remain in operation (dark green scenario), facility 1 has a 51% probability of being the most patronized, while facility 2 has a 48% probability of being so. Similarly when all facilities remain open (yellow scenario), facility 1 has a 48% probability of being the most used, while facility 2 and 3 have, respectively, a 39% and a 14%. Note that in this scenario, there

are some cases where adding up the three probabilities yields a result greater than 100% (e.g. case I). This is because in some iterations two facilities are tied for the largest market share.

- The median of the number of customers arriving at each facility. For instance, when facilities 1 and 2 remain open, the median arrival rates are respectively 61 and 59 arriving customers.
- The letters (i.e. “a”, “b”, “c”, and “d”) next to the median value of the second facility of each scenario indicate the results of the MWW-test for the null hypothesis that the median arrival rate is the same for the assessed facilities. The letters “a” and “b” indicate that this hypothesis is rejected at a significance level of 0.01 and 0.1, respectively, against the alternative hypothesis that the median arrival rate of the most conservative manager’s facility is significantly greater than that of the other facility. The letter “c” indicates that the null-hypothesis is rejected at a significance level of 0.01, against the alternative hypothesis that the median arrival rate of the most conservative manager’s facility is significantly lower than that of the other facility. The letter “d” indicates that the null hypothesis cannot be tested because of lack of data. For instance, the letter “a” in the scenario where facility 1 ($\delta 1$ = 0.8) and 3 ($\delta 3$ = 0.2) remain open (medium green) indicates that according to the MWW-test the median number of customers using facility 1 (i.e. with the most conservative manager) is significantly greater than that using facility 3 at a significance level of 0.01. In the scenario where the three facilities remain open, this test is assessed by pairs of facilities and the significance of the test is indicated in the last three columns for each pair of facilities. For instance, in case I, the MWW-test indicates that at a significance level of 0.01, the median of facility 1 ($\delta 1$ = 0.8) is significantly greater than that of facility 2 ($\delta 2$ = 0.5) and facility 3 ($\delta 3$ = 0.2) and that the median of facility 2 ($\delta 2$ = 0.5) is significantly greater than that of facility 3 ($\delta 3$ = 0.2).

In most cases, the p-values computed by MWW-test suggest rejecting the null hypotheses at the 1% level of significance. This enables us to conclude that the facility of the most conservative manager (i.e. manager 1, $\delta 1$ = 0.8) usually attracts more customers than the other facilities. Most of the exceptions are due to lack of data. For instance, the yellow scenario (i.e. the three facilities remain open) in case IX (ψ = 0.8 and ζ = 0.8), which is unlikely (0.2% of iterations barely visible in Figure 3.8).

				2 FACILITIES OPEN						3 FACILITIES OPEN					
Case	ζ	ψ	Measure	F1 and F2		F1 and F3		F2 and F3		All facilities			F1-F2	F1-F3	F2-F3
				F1	F2	F1	F3	F2	F3	F1	F2	F3			
I	0.2	0.2	P(equal)*	0.85%		0.76%		2.94%		-					
			P[λ(Fx)=Max(λ)]	51%	48%	75%	24%	84%	13%	48%	39%	14%			
			Median	61	59	69	51 a	70	50 a	48	42	28	a	a	a
II	0.2	0.5	P(equal)*	2.36%		0.70%		-		-					
			P[λ(Fx)=Max(λ)]	57%	41%	77%	22%	88%	12%	54%	36%	10%			
			Median	63	57 a	69	51 a	72	48 a	51	42	27	a	a	a
III	0.2	0.8	P(equal)*	1.12%		1.30%		3.51%		-					
			P[λ(Fx)=Max(λ)]	56%	43%	79%	19%	86%	11%	58%	33%	10%			
			Median	64	56 a	72	48 a	71	49 a	52	41	26	a	a	a
IV	0.5	0.2	P(equal)*	1.58%		0.68%		-		-					
			P[λ(Fx)=Max(λ)]	54%	44%	78%	21%	100%		47%	38%	15%			
			Median	63	57 a	79	41 a	88	32 a	50	41	26	a	a	a
V	0.5	0.5	P(equal)*	1.03%		-		-		-					
			P[λ(Fx)=Max(λ)]	67%	32%	88%	12%	89%	11%	57%	37%	7%			
			Median	73	47 a	87	33 a	82	38 a	53	41	22	a	a	a
VI	0.5	0.8	P(equal)*	0.86%		-		-		-					
			P[λ(Fx)=Max(λ)]	76%	23%	100%	-	100%	-	62%	25%	13%			
			Median	80	40 a	84	36 a	87	33 a	57	35	21	a	a	a
VII	0.8	0.2	P(equal)*	1.42%		2.25%		-		-					
			P[λ(Fx)=Max(λ)]	39%	60%	56%	42%	86%	14%	-	43%	57%			
			Median	52	68 c	63	57	74	46 a	14	55	60		b	
VIII	0.8	0.5	P(equal)*	0.79%		2.13%		-		-					
			P[λ(Fx)=Max(λ)]	66%	33%	72%	26%	100%	-	11%	67%	22%			
			Median	69	51 a	105	15 a	104	16 d	54	18	48	a		
IX	0.8	0.8	P(equal)*	0.51%		-		-		-					
			P[λ(Fx)=Max(λ)]	82%	18%	100%	-	-	-	100%	-	-			
			Median	85	35 a	112	8 a	-	-	62	40	18	d	d	d

* P(equal) represents the relative frequency of iterations in which the number of arriving customers is the same for all facilities open in each scenario.

H_0: The median of the number of arriving customer is the same for all the facilities open;
against:

(i) H_a: The median of the number of customers arriving at the most conservative manager's facility is significantly *greater* than that of the number of customers arriving at the other facility. According to the MWW test, H_0 can be rejected at significance levels of: [a] 0.01; [b] 0.1;

or

(ii) H_a: The median of the number of customers arriving at the most conservative manager's facility is significantly *lower* than that of the number of customers arriving at the other facility, H_0 can be rejected at a significance level of: [c] 0.01.

[d] the hypothesis cannot be tested because of lack of data: These are scenarios which occur very rarely (see Figure 3.8).

Table 3.5 Some statistics for the number of customers arriving at each facility during steady-state for 1,000 simulations of the model as a function of ψ and ζ

In case VII, where managers are moderately irrational (ψ = 0.2) and fast decision makers (ζ = 0.8), the scenario in which the most reactive manager (i.e. manager 3, $\delta 3$ = 0.2), shuts down his facility (i.e. dark green scenario) contrasts with the expected behaviour. That is, the less conservative manager of the two still active managers, who is manager 2 (i.e. $\delta 2$ = 0.5), usually captures the largest market share.

According to Table 3.5, the scenarios in which all remaining facilities serve on average the same number of customers in steady state have very little chance of occurring, no matter how irrational or how fast the managers are (see P(equal)). It is worth mentioning that this scenario does not necessarily imply customers being equally split between all the facilities. A configuration with groups of alternating customers is another possibility which can yield this scenario.

Table 3.6 contain the minimum, maximum and average closure times of facilities depending on which and how many facilities close, for the same cases discussed above. Recall that for these cases, the scenario where the three facility close is unusual. For those scenarios where only one facility closes, we also compute aggregate values (see Aggregate in Table 3.6): the minimum and the maximum closure times are respectively the lowest minimum value and the highest maximum value between the three possible scenarios, while the average is the weighted average of the three scenarios. For instance, in case 1 when only one facility closes, this occurs between periods 19 and 477, on average in period 46. When two facilities close, the minimum, maximum and average closure times are given according to the order in which these facilities closed, no matter which facility close first. For instance, in case 1, when two facilities close, the first does so anywhere between periods 21 and 193, on average in period 47, and the second one between periods 25 and 193, on average in period 55.

We can observe that for all cases the most likely scenarios (recall Figure 3.8) have highly asymmetric distributions of closure time of the facilities. The exceptions are the scenarios where two facilities close in cases II and III and the one where facility 1 is the only closing in case VIII. These scenarios have a low probability of occurrence as can be seen in Figure 3.8.

Note that, except for case I, the average closure time of the first facility closing in the scenario where two facilities do so is lower than when only one facility closes. Thus, we can say that in these cases the first closure of a facility occurs, on average, earlier in the scenarios when two facilities close than in those where only one facility shuts down.

				1 FACILITY CLOSING				**2 FACILITIES CLOSING**	
Case	ζ	ψ	**Measure**	**F1 Closing**	**F2 Closing**	**F3 Closing**	**Aggregate**	**First Closing**	**Second Closing**
I	0.2	0.2	Min	31	21	19	19	21	25
			Average	35	44	51	46	47	55
			Max	47	408	477	477	193	193
II	0.2	0.5	Min	31	22	20	20	25	26
			Average	44	46	57	52	28	30
			Max	401	499	492	499	30	33
III	0.2	0.8	Min	32	22	20	20	23	26
			Average	36	52	71	61	29	36
			Max	53	490	464	590	35	44
IV	0.5	0.2	Min	28	16	13	13	13	17
			Average	53	85	80	80	31	44
			Max	384	395	467	467	306	315
V	0.5	0.5	Min	28	17	13	13	13	18
			Average	34	106	96	91	39	60
			Max	51	488	483	488	384	472
VI	0.5	0.8	Min	28	18	13	13	14	19
			Average	32	87	110	93	80	175
			Max	38	485	488	488	462	497
VII	0.8	0.2	Min	28	14	12	12	12	15
			Average	164	38	40	43	31	111
			Max	440	257	330	440	392	489
VIII	0.8	0.5	Min	32	16	12	12	12	17
			Average	101	90	51	55	47	172
			Max	169	446	348	348	379	498
IX	0.8	0.8	Min	-	24	9	9	8	17
			Average		167	38	47	44	173
			Max	-	290	328	328	322	496

Table 3.6 Minimum, maximum and average closure time of facilities for 1,000 simulations of the model as a function of ψ and ζ

Recall that when managers take decisions slowly (cases I, II and III), the scenarios of two facilities closing are very rare (recall blue shades in Figure 3.8). Conversely, these scenarios are more likely in those cases where managers take decisions quickly (cases VII, VIII and IX). In Table 3.6 we can observe that when these scenarios occur, the second facility usually shuts down soon after the first did so for those cases where managers are slow decision makers (cases I, II and III). In contrast, the second facility takes generally significantly longer to close, when manager are fast decision makers (cases VII, VIII and IX).

In those cases where managers are either slow or moderately fast decision makers (cases I to VI) and only one facility close, facility 1 (whose manager is the most conservative $\delta 1$ = 0.8) is the least likely to close, but when it does, its average closure time is the lowest.

The main conclusion we can extract from the above experiment is the fact that the most conservative managers capture the largest market share and that their facilities are less likely to shut down. In order to validate whether this conclusion also applies for different allocations of the three kinds of managers around the three facilities, this experiment was also performed for other combinations of the sequence {0.2, 0.5, 0.8} of the managers' coefficients of expectations. That is, we move the different managers (i.e. reactive, hesitant and conservative) around the three positions in the model (facilities 1, 2 and 3). The results we obtain do not differ much from those presented above and yield similar conclusions about the behaviour of reactive and conservative managers.

We repeated the experiment with the same combinations of the managers' coefficient of expectations for another combination of the customers' parameters. We have tested all the extreme and intermediate cases regarding the customers' attitudes toward the new information. Again, the main insights about the market share and the probability of the conservative manager's facility remain valid. One significant observation is that the scenario in which the three facilities are still open at time 500 is much more frequent when customers are either hesitant or very conservative regarding new information, no matter its provenance. Conversely, the scenario where the conservative managers achieve a monopoly position is more likely when customers are more reactive with respect to their own information and more conservative regarding that of their neighbours.

3.4.2.3 Influence of the benchmark on the average sojourn time of the system.

Next we study the impact of the exogenous market reference, which the managers use as benchmark to adjust their service capacity, on the facilities' performance and the aggregated performance of the system. We have run 1,000 iterations of the model for thirty-eight different values of the reference average sojourn time (τ_{MR}) (ranging from 0.3 to 4.0) and varying also the parameters ψ and ζ. All the other model parameters remain the same as in Table 3.3: $\delta 1$ = 0.8, $\delta 2$ = 0.5, $\delta 3$ = 0.2, α = 0.3 and β = 0.7. Each simulation is run for 500 time periods and the weighted average sojourn time in steady state is computed for the last 40 time period. All the combination of parameters ψ and ζ yield similar distributions and tendencies to those shown in Figure 3.9 which corresponds to the case with ψ = 0.8 and ζ = 0.5.

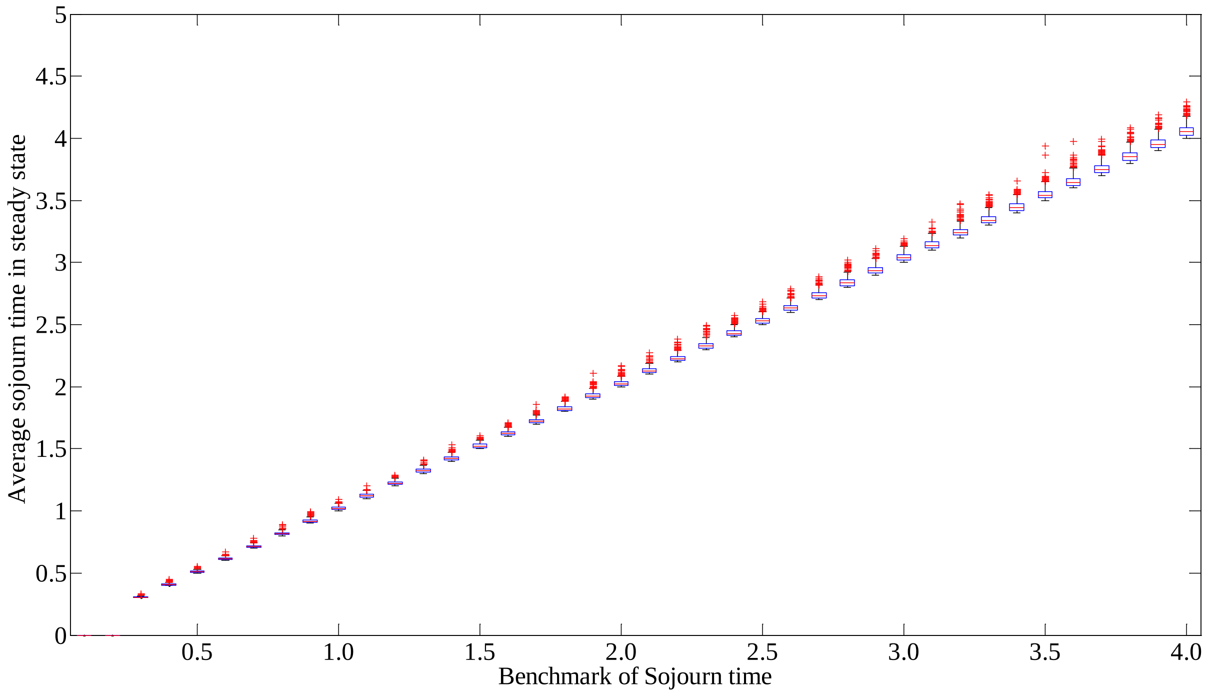


Figure 3.9 Evolution of the distribution of average sojourn time of the system in steady-state generated through 1,000 simulations of the model as a function of the benchmark of sojourn time (τ_{MR}). Managers and customers are allocated with different initial expectations and the following parameters: $\delta 1$ = 0.8 , $\delta 2$ = 0.5 , $\delta 3$ = 0.2 , ψ = 0.8 and ζ = 0.5 w.r.t managers and α = 0.3 and β = 0.7 w.r.t. customers.

When analysing the typical behaviours exhibited by the system and its agents, we mentioned that the benchmark of the average sojourn time is an attractor of the system since the interaction between managers and customers causes the weighted average sojourn time of the system to converge to this benchmark. Figure 3.9 enables us to generalise this claim. In this figure we can observe that the weighted average sojourn times of the system in steady state are closely clustered around the benchmark, whatever the value of this benchmark. This means that for a given value of the benchmark, the average sojourn time converges to this value. The managers' and customers' profile only influence the time the system takes to converge.

The small variations in average sojourn time visible in Figure 3.9 are precisely due to this characteristic of convergence and to the low probability of long transient periods. The higher the benchmark, the greater the dispersion of the weighted average sojourn times.

3.5 Conclusions

In this chapter we have extended the CA model proposed by Delgado et al. (2011a) and Sankaranarayanan et al. (2011) by incorporating the service rate as an endogenous variable. That is, we have endowed the managers with the ability to adjust the service capacity of their facility. In this sense we introduce additional parameters into the model which characterise the managers' profile and define the decision rule they apply to adjust the service capacity. Managers are also provided with a memory which enables them to update their expectations regarding the number of customers arriving at their facilities each period. For this purpose, coefficients of expectations ($\delta 1$, $\delta 2$, $\delta 3$) are allocated to managers. Additionally, we have characterised the managers depending on the extent to which they account for their previous decisions when deciding by how much to adjust capacity and the speed at which they take decisions. The first attribute indicates the level of irrationality of managers when the decision making involves delays. With the purpose of measuring this level we have introduced a parameter called "coherence factor" (ψ), which captures the proportion of the managers' previous decisions, which have not been yet implemented, that they consider when making decisions. Concerning the other attribute, i.e. the speed, we incorporate another parameter to capture the proportion of the desired adjustment they decide to implement (ζ).

In order to study the collective behaviour which emerges from the interaction between customers and managers over time, we have run the simulation model for different configurations of parameters using the same initial conditions. We have also performed some experiments to analyse the sensitivity of the model to the different parameters involved in the system. We conclude that the managers' decisions are strongly "path-dependent", in the sense that two or three managers can have the same profile, but behave very differently from each other due to the initial memories allocated to managers and customers.

Another conclusion is that the facility of the most conservative manager usually achieves the highest market share. Additionally this facility is the most likely to remain open until the end of the simulation period (i.e. it is less likely that this facility shuts down).

The analysis of the typical collective behavioural patterns which emerge when the system reaches steady-state and the third experiment which concerns the benchmark of sojourn time enable us to conclude that this benchmark is an attractor point of the system. This means that the average sojourn time of the system will converge to the benchmark over time.

Nevertheless, the rate at which this convergence occurs depends on the managers and the customers’ profiles. In some cases, the convergence period can be very long.

This work can be extended by analysing the sensitivity of the model to the customers’ parameters and the delays involved in the implementation of the managers’ decisions. Further work in this field includes adding uncertainty parameters to the customers and managers’ decision rules (as was done in the previous chapter for the customers) and assessing the influence of other service factors in the customers and managers’ decisions, such as price and quality.

4 MANAGING CAPACITY AT A SERVICE FACILITY: AN EXPERIMENTAL APPROACH

ABSTRACT

In this chapter, we perform an experiment through which we address the capacity management of a service facility. This experiment was carried out using a computational user interface based on a system dynamics model. The model represents a queueing system with one facility and an infinite number of virtual customers who decide whether or not to patronise the facility. The facility has a manager whose role is played by human subjects. The subjects were recruited among undergraduate and master students in Finance, Management and Economics at the University of Lausanne. Their goal was to maximise the profits of the service facility by adjusting the service capacity. The capacity adjustment process implies certain implementation times. These are the capacity delivery and dismantling delays which apply when subjects decide either to add or withdraw capacity, respectively. There are also delays due to the time customers take to update their perceptions. Subjects were aware of the delays involved in the capacity adjustment process, but not of those concerning customers. Additionally, the subjects were provided with information regarding the number of customers waiting for service (i.e. backlog of work), the available service capacity, the utilisation rate of the service facility and all the information regarding cost, revenue and profits per period as well as the cumulative profits. We use the results to study the way subjects in a laboratory environment face a situation in which they must adjust the capacity of a service facility considering that customers do not like to wait for service. The results show that the subjects can be classified as belonging to one of three types of managers labelled as: gradual, lumpy and random investors. Our analysis also shows that as the delivery and dismantling delays increase, the cumulative profits achieved by the subjects decrease significantly.

KEYWORDS: Queueing systems, capacity adjustment management, system dynamics (SD), experimental economics, potential customers and customer base.

4.1 Introduction

In chapter 3, we used CA to model the way managers of a service facility adjust its capacity based on their expectations regarding the future customer arrivals. These managers knew the number of customers arriving at their facilities and used this information to update their expectations. Now we use a SD model, which was proposed by van Ackere et al. (2010) and subsequently adapted by Delgado et al. (2011c) (see Appendix D), as an experimental platform to collect information about how subjects playing the role of a manager in a laboratory environment, manage the capacity of a service facility. We build a computational graphical user interface through which subjects take decisions to adjust the capacity of a garage (i.e. the service facility) each month (i.e. the time unit)

We apply the protocols of experimental economics (Friedman and Sunder 1994) to perform the experiment and control the exogenous factors which could bias the results. Experimental economics is a methodology applied for collecting data from human subjects and studying their behaviour in a controlled economic environment (Friedman and Sunder 1994). In behavioural economics SD models have been widely used to simulate socio-economic environments. Some examples include Sterman (1989a and 1989b), Kampmann (1992) and Moxnes (2000). For more details about how SD models are used to carry out laboratory experiments, see Arango et al. (2012).

Few experiments in economics and management science study queueing problems. Rapoport et al. (2004) formulated a queueing problem with endogenously determined arrival rates and state-dependent feedback as a non-cooperative n-person game. These experiments are focused on studying behaviour from the point of view of the customers. Subjects, playing the role of car owners who need to take their car to a garage for the emissions control, should decide each experimental trial whether or not to join the queue. If they choose to join it, they also should decide when to do so. The authors first analyse the strategies which lead to the equilibrium solutions for the game. Then, they perform an experimental study in order to identify if customers, making individual decisions, achieve a coordinated solution for this game in large groups. Finally, they characterise this solution and determine if it converges to the equilibrium. Each player was provided with information about his arrival time, his payoff for the trial, his cumulative payoff, his waiting time and the number of players who arrived at the same time for service. Players were not provided with information about the other members of the group. Subsequently Seale et al. (2005) extended this work to non-cooperative n-person games with complete information (i.e. including the information of the other group-members). Later Stein et al. (2007) and Rapoport et al. (2010) performed other experiments to study queueing systems with endogenous arrival rates and batch service. They analyse how customers decide whether to join a queue and when to do so.

In this chapter we focus on the manager's role. We study the capacity management of a service facility in which the managers' decisions take time to be implemented and customers take time to update their perceptions. Our aim is to analyse how subjects playing the role of managers make decisions regarding the capacity of a service facility in which customers must wait to be served and the customers' satisfaction depends on this wait. Additionally we want to study the effects of the delays involved in the system on the managers' decisions. The problem is portrayed as the management of a garage where customers must make an appointment to take their cars. The customers' wait runs from the time they make their appointment to the time their car is serviced. The task of the subjects is therefore to manage the capacity (i.e. add or remove capacity) of the garage in order to satisfy their current customers and attract potential customers. When subjects decide to increase capacity, the capacity orders materialise after a delivery delay. Similarly, the retirement decisions involve a dismantling delay. We hypothesise that, when making their decisions, subjects only partially account for past decisions which have not yet been implemented.

Our results indicate that the subjects can be classified into three types of managers according to the way in which they adjust the service capacity: those who make gradual investments in service capacity; those who make lumpy investments; and those who overreact to the customer behaviour by adding and removing capacity without any logic. The autocorrelation analysis of the capacity adjustments, which subjects made each period, shows that these adjustments strongly depend on the decisions taken one period before. In order to determine the way subjects adjust service capacity, we propose a decision rule which estimates the capacity adjustment decision using a multivariate linear regression in terms of the current backlog, the currently available service capacity and past decisions which have not yet implemented. The statistical tests for the significance of the estimated parameters indicate that the backlog and the available service capacity influence the capacity adjustment decisions. Finally a statistical comparison between treatments, in which the length of the delays involved in the system is varied, indicates that the delivery and dismantling delays have a significant impact on the cumulative profits achieved by the subjects.

Some of the results obtained from this experiment have been published in Delgado et al. (2011c).

The remaining sections of this chapter are organised as follows: In the next section, we briefly describe the SD model used in this chapter focusing on the adaptations made to the original model developed by van Ackere et al. (2010). In section 4.3 we explain the way we carry out the experiment following the protocols of experimental economics (Friedman and Sunder 1994). In the last section we discuss the results. We first provide a descriptive analysis of the way subjects take their decisions. Next, we test the experimental hypothesis and analyse the difference treatments.

4.2 A Service Facility Management Model

The queueing model used in this chapter was originally proposed by van Ackere et al. (2010) where a full explanation of the model can be found. We use this model as a computational platform to perform a laboratory experiment with human subjects. Thus, we will explain how we have adapted this model for the experiment.

van Ackere et al. (2010) use SD to model the feedback and delay structure involved in the relationship between customers and the manager of a service facility. Customers decide whether to join or not the facility for service, while the manager adjusts the service capacity in order to attract more customers. This structure is similar to the one explained in the previous chapter (see Figure 3.1). Nevertheless, the structure of the system differs in two main aspects: 1) there is only one facility; and 2) two groups of customers are assumed: current and potential customers. The number of customers queueing is modelled as the backlog of work which the facility has to serve over the following periods. Likewise, the assumption of a captive market (i.e. customers have to join the facility) is also relaxed in this chapter by modelling customers who can decide whether or not to use the facility for service.

Another difference with regard to the previous chapters is the sojourn time which customers experience. In the previous chapters we estimated an average sojourn time of customers at the facilities in a transient state (i.e. arrival rates might momentarily be larger than the service rates) using a non-linear equation in terms of the arrival and service rates. Now, given the continuous nature of SD models, the estimate of the sojourn time is simpler.

This sojourn time is considered to be the time between the moment customers make an appointment for service and the moment their service is completed. Once the customers make their appointment they become part of the backlog of customers waiting for service. Thus, the sojourn time of these customers depends on the service rate (μ_t) of the facility. We compute the sojourn time (W_t) each period as the ratio between the backlog of customers (i.e. queue Q_t) and the service rate (μ_t):

$$W_t = \frac{Q_t}{\mu_t} \tag{4.1}$$

The service rate depends on the service time and the available service capacity. The service time is assumed to be exogenous and fixed, while the available service capacity is endogenous and depends on the manager's decisions. The latter is explained below.

Current customers make up the customer base of the facility; they periodically patronise it as long as they are satisfied. Their level of satisfaction depends on how their expected sojourn time ($\dot{W}_t$) compares to a market reference (τ_{MR}). This market reference has the same interpretation as the benchmark used in the previous chapter, but now it also applies to the customers. All current customers deciding to join the facility at given time *t* base this decision on the same expected sojourn time ($\dot{W}_t$). This expectation is updated using the sojourn time (W_t) experienced by the customers served each period through the following equation:

$$\dot{W}_t = \varphi * W_{t-1} + (1-\varphi) * \dot{W}_{t-1} \tag{4.2}$$

where $1/\varphi$ is assumed to be the time taken by customers to adapt their expectations. So current customers compare their expected sojourn time to the market reference and decide whether or not to stay with this facility.

Potential customers are the prospects for the facility. That is, those customers who the manager envisages as potentially attractive to the business. They can be either former customers, who left the facility due to dissatisfaction, or new customers who require the service and are looking for a facility. Potential customers decide whether or not to become current customers of the facility depending on their perceptions about the sojourn time. They form their perceptions through the word of mouth effect using Equation (4.2). Updating perceptions based on the reputation of a certain facility often requires more time than when based on one's own experience. Indeed, the information about this facility is less readily available to potential customers than to current customers. Thus, we assume that the time required by potential customers to adapt their expectations is longer than or equal to that of the current customers.

While the current customers' expectations determine their loyalty to the facility, the potential customers' perception defines whether they will try the facility. The lower the sojourn time expected by current customers, the more loyal they are, whereas the higher this expectation the more customers will leave the customer base. Regarding potential customers, the lower their perceptions, the more they will try out the facility. The rates at which new customers join and current customers leave are modelled using nonlinear functions of the satisfaction level. van Ackere et al. (2010) discuss some alternatives to model these functions.

The manager can adjust the service capacity of his facility whenever he wishes, but this adjustment involves an implementation time (e.g. delivery and retirement delays). van Ackere et al. (2010) represent this time using an information delay (Sterman 2000): after the manager estimates the required capacity, any needed adjustment is implemented gradually. This is a simplified view of the delay structure. In a SD context, this kind of delays is better modelled through material delays, which capture the real physical flow of the capacity (Sterman 2000). Once the adjustment decision has been made, its implementation is not immediate. We deviate from van Ackere et al. (2010) by incorporating this material delay structure in the model, using the stock and flow diagram of Figure 4.1. In this way, we can model how the manager accounts for his previous decisions, which have not yet been implemented, when taking his next decision.

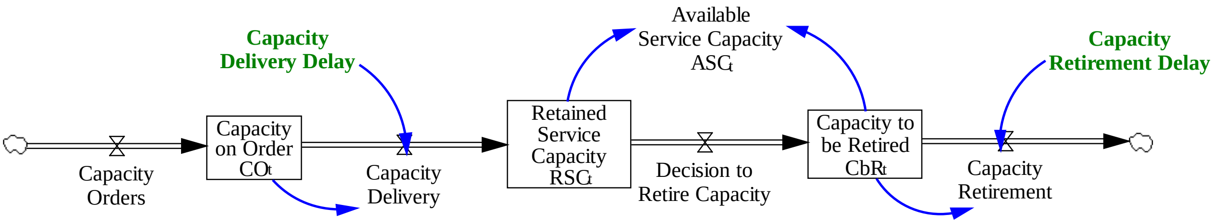


Figure 4.1. Stock and flow diagram for the capacity adjustment management of a service facility.

The capacity adjustment process is depicted in Figure 4.1 by capacity orders and retirement decisions. On the one hand, when the manager decides to add capacity (c.f. capacity orders in Figure 4.1), these orders accumulate as capacity on order (CO_t) until they are available for delivery. After a capacity delivery delay, the capacity on order is delivered at the facility and placed as retained service capacity (RSC_t).

On the other hand, the capacity retirement decisions also take time to be carried out. In this sense, once the manager decides to withdraw capacity, this capacity is removed from the retained service capacity and earmarked as capacity to be retired (CbR_t), as shown in Figure 4.1. This capacity remains available for service until the capacity retirement delay expires. Once this occurs, this capacity is effectively retired from the facility. Thus, the available service capacity (ASC_t) at the facility at time t is made up of the retained service capacity (RSC_t) and the capacity to be retired (CbR_t). This relation is shown in Figure 4.1 and given by the following equation:

$$ASC_t = RSC_t + CbR_t \tag{4.3}$$

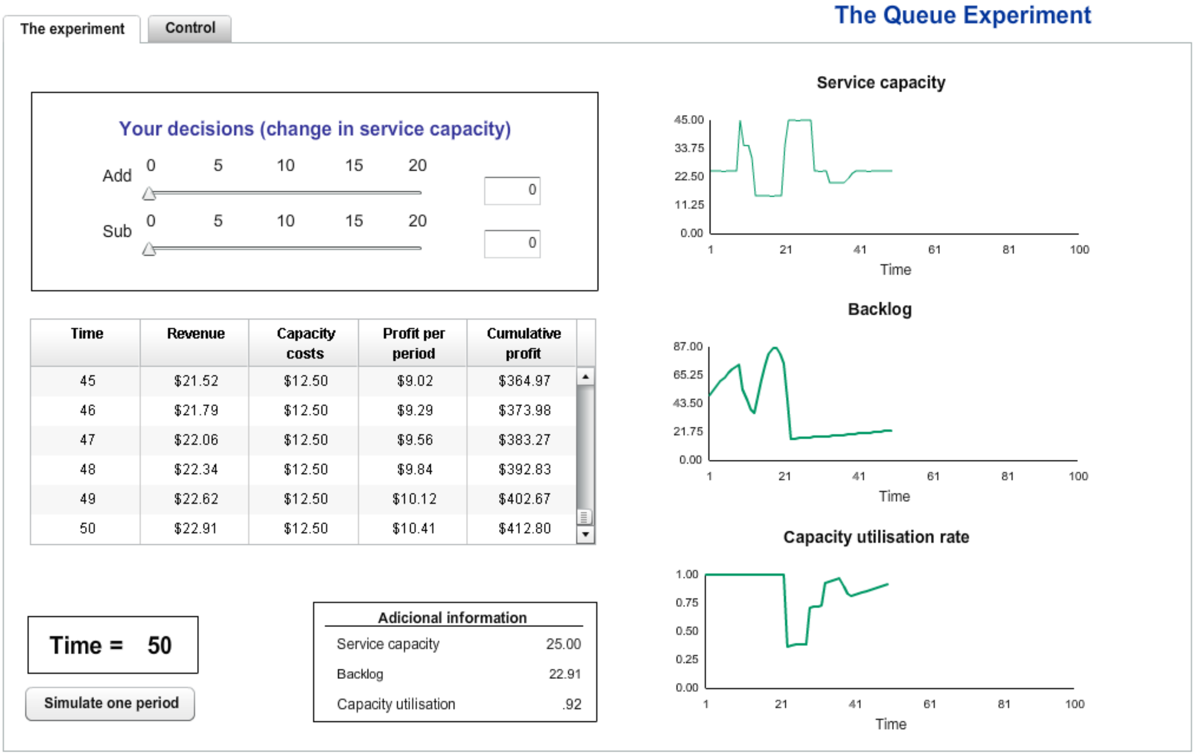
As the available service capacity (ASC_t) increases, so does the service rate. Consequently the backlog of work will decrease more quickly and the customers' wait will be lower. This results in less customers leaving and more prospects joining. The opposite logic applies when the available service capacity (ASC_t) decreases.

The behaviour generated by the model in equilibrium has been discussed in Delgado et al. (2011c). In that paper the authors proposed two decisions rules to model the capacity adjustment decisions. The system behaviour generated by these two approaches was compared to the one observed in the laboratory. Here we focus on analysing how the subjects use the available information about the service facility and customer behaviour when making capacity adjustment decisions. We propose another decision rule and analyse statistically whether it accurately captures the subjects' decision process.

4.3 A Service Facility Management Experiment

The experiment reported in this section was developed to collect information about the way human subjects, taking on the role of a manager, face a situation in which they must adjust the capacity of a service facility. This experiment addresses the capacity management of a service facility, where a certain delay between applying for the service (making an appointment) and receiving it, is considered normal. Examples include car maintenance (as in this experiment) and hairdressers. This experiment is based on the simulation model described above and performed using a computational graphical interface developed using the Forio Simulation Platform (Forio Online Simulations, http://forio.com/simulate). Forio Simulation is an online platform which enables us to easily create graphical user interfaces based on simulation models.

The experiment portrays the management of a large garage for repair and maintenance of cars. The user interface is shown in Figure 4.2. This is divided in two parts. The first part is the input panel where the subjects indicate their decisions regarding the service capacity. And the second part shows the information of the garage which is given to the subjects. This information is updated each time they take a decision, as shown in Figure 4.2 for a subject who is half-way through the experiment (time=50).



Your decisions (change in service capacity)

Add 0 5 10 15 20 0

Sub 0 5 10 15 20 0

Time	Revenue	Capacity costs	Profit per period	Cumulative profit
45	$21.52	$12.50	$9.02	$364.97
46	$21.79	$12.50	$9.29	$373.98
47	$22.06	$12.50	$9.56	$383.27
48	$22.34	$12.50	$9.84	$392.83
49	$22.62	$12.50	$10.12	$402.67
50	$22.91	$12.50	$10.41	$412.80

Time = 50

Simulate one period

Adicional information

Service capacity	25.00
Backlog	22.91
Capacity utilisation	.92

Figure 4.2 Experimental interface for the service capacity adjustment of a service facility

The garage is configured with an existing customer base as well as an infinite number of potential customers who are not currently using the garage, but might consider doing so in the future. In period 1 there is a backlog of 50 cars and the garage is endowed with an initial capacity to serve 25 cars simultaneously. Each car required minimum 1 month to be serviced (i.e. the time between the appointment and the completion of service). Servicing a car yields a revenue of 1$. Service capacity entails a fixed cost of 0.5$ per month independently of whether or not it is being used. Table 4.1 contains the other relevant parameters and initial conditions. Note that although SD models are continuous, as mentioned above, we perform this experiment using discrete time periods. Each experimental period represents one month in the problem context.

The experimental design considers five treatments where we vary the different delays implied in the decision process. We want to test if these delays have an impact on the profits the manager achieves. We hypothesise that the delay structure, inherent to the system, affects how the manager decides to adjust capacity. This delay structure includes the delays the manager knows (i.e. the lags in capacity adjustment), and those which are unknown to him (i.e. the time required by potential and current customers to update their perceptions). Table 4.2 summarises the conditions of each treatment.

State Variables	**Value**	**Unit**
Initial Customer base	175	People
Initial Backlog (i.e. Queue)	50	People
Capacity on order	0	People / month
Initial Service capacity	25	People / month
Capacity to be retired	0	People / month
Perceived waiting time of current customers	2	Months
Perceived waiting time of potential customers	2	Months
Exogenous Variables	**Value**	**Unit**
Service frequency	0.15	1 / month
Market reference waiting time (τ_{MR})	2	Months
Delays	**Value**	**Unit**
Capacity delivery delay	4	Months
Capacity retirement delay	2	Months
Perception time of current customers (1 / φ_c)	2	Months
Perception time of potential customers (1 / φ_p)	4	Months

Table 4.1 Initial configuration of the model used to represent the management of a garage

The purpose of treatments A and B is to study the impact of the length of delays about which the manager does not know when taking decisions about the service capacity, i.e. the time customers take to update their expectations about the garage. Treatments C and D address those delays of which the manager is aware, i.e. the time required to implement his capacity adjustment decisions.

Treatment	Current customers Delay	Potential customers Delay	Time to increase capacity	Time to decrease capacity	Number of subjects*
Base Case	4	2	4	2	33 (3)
A	10	2	4	2	31 (1)
B	6	4	4	2	32 (3)
C	4	2	8	4	31 (5)
D	4	2	2	1	30 (1)

* The first number is the total number of subject assigned to the treatment, while the number in parenthesis is the number of subjects who closed down the facility before period 100

Table 4.2 Treatment conditions.

4.3.1 Experimental Protocol

We have designed this experiment based on the protocols from experimental economics (e.g. Smith 1982, Friedman and Sunder 1994). This experiment was conducted in the informatics laboratories of the School of Business and Economics of the University of Lausanne. Subjects were allocated across eleven experimental sessions according to their availability. Each session involved around sixteen subjects and lasted on average 2 hours, including the time for preparing the lab and collecting results. Two facilitators supervised each session.

Upon arrival at the laboratory, the subjects were allocated to a PC and separated from their neighbours by another PC. Communication between the subjects was forbidden. Once they were seated, we provided them with written instructions and a consent form, which they had to sign before starting the experiment. The instructions were quite simple and provided subjects with a short explanation of the system that they had to manage in the experiment and with all the information which they had available to carry out their task. These instructions can be found in the appendix of Delgado et al. (2011c) (see Appendix D).

Before starting the experiment, a short introduction was presented to the subjects with the aid of a Power-Point presentation. We explained the interface and the tasks they should perform during and after the experiment.

4.3.1.1 Subjects

The sample is drawn from undergraduate and master students in Finance, Management and Economics from the University of Lausanne. We invited subjects to participate in an experiment designed to study decision making in a service industry. The students were motivated to participate in the experiment by the possibility to earn up to 80 Swiss Francs, depending on their performance. We received about four-hundred replies and recruited one-hundred and fifty-seven subjects following the principle of "first come, first served". The subjects were allocated across the five experimental treatments attempting to keep homogenous samples according to the participants' profile. Each treatment had at least thirty participants.

4.3.1.2 The subjects' task

The subjects' objective was to maximise the cumulative profits of the garage over 100 experimental periods (i.e. months). Their task was therefore to manage the service capacity of the garage to satisfy their existing customers and attract new ones. Subjects should decide each period whether to adjust or not the service capacity and by how much. The capacity adjustment implied either to add or remove capacity. When subjects decided to adjust capacity they could use the sliders in the first square (Your decisions) of Figure 4.2 or write their decision in text box on the right side of the sliders. Subjects were informed that their decisions would not materialise immediately. They were aware of the delivery and dismantling delays of the treatment (see Table 4.2) they had been assigned to. The experiment ended once the subject made the last decision in period 100 or when the available service capacity reached 0, i.e. the subject shut down his facility. Subjects who shut down the facility before time 100, will not be considered in the statistical analysis and when comparing different treatments

Subjects were asked to register each decision in writing to avoid loss of data in case of trouble during the experiment (e.g. a power cut, an error on the online platform, among others).

4.3.1.3 The available information

To help subjects make these decisions we provided them with information about the system (see Figure 4.2). Subjects had information about the number of customers currently waiting for service or whose car was currently being serviced (referred to as the backlog). They knew the current service capacity and the utilisation rate of this capacity. This information was also given in a graphical way so that they could observe its evolution. Additionally they could observe the current revenue, the capacity costs, the profit per period and the cumulative profit.

4.3.1.4 Monetary rewards

At the end of the experiment subjects received a reward which consisted of two parts. If they pursued the experiment until the end, they received a guaranteed participation fee of 20 CHF. Additionally, depending on their performance, subjects could receive a bonus which varied between 0 and 60CHF. The performance was measured through the total profit they achieved at the end of the experiment. Subjects knew beforehand the payoff scale linking their profits to their bonus.

4.4 Experimental Results

Based on the results we identify three groups of subjects whose decisions yield similar behavioural patterns. In the base case, thirteen subjects are classified as group 1, nine subjects as group 2 and eleven subjects as group 3. Figure 4.3 illustrates the capacity adjustment decisions of a typical subject of each group as well as the evolution of the backlog and the available service capacity over the 100 periods. Negative capacity adjustment decisions represent capacity retirement decisions.

During the first six periods the backlog increases. This rise is independent of subjects' decisions and occurs because of the initial conditions which imply a 2 month sojourn time for the initial backlog of customers. This corresponds to the market reference and the initial perception of both current and potential customers. Hence, this configuration attracts new customers to the facility for the first periods. As the service capacity remains constant during these periods (capacity orders placed in period 1 are not fulfilled until period 5), the backlog of customers rises. Consequently, the sojourn time also increases. This affects the customers' perceptions, resulting in a drop in the backlog from period 6, as shown in Figure 4.3.

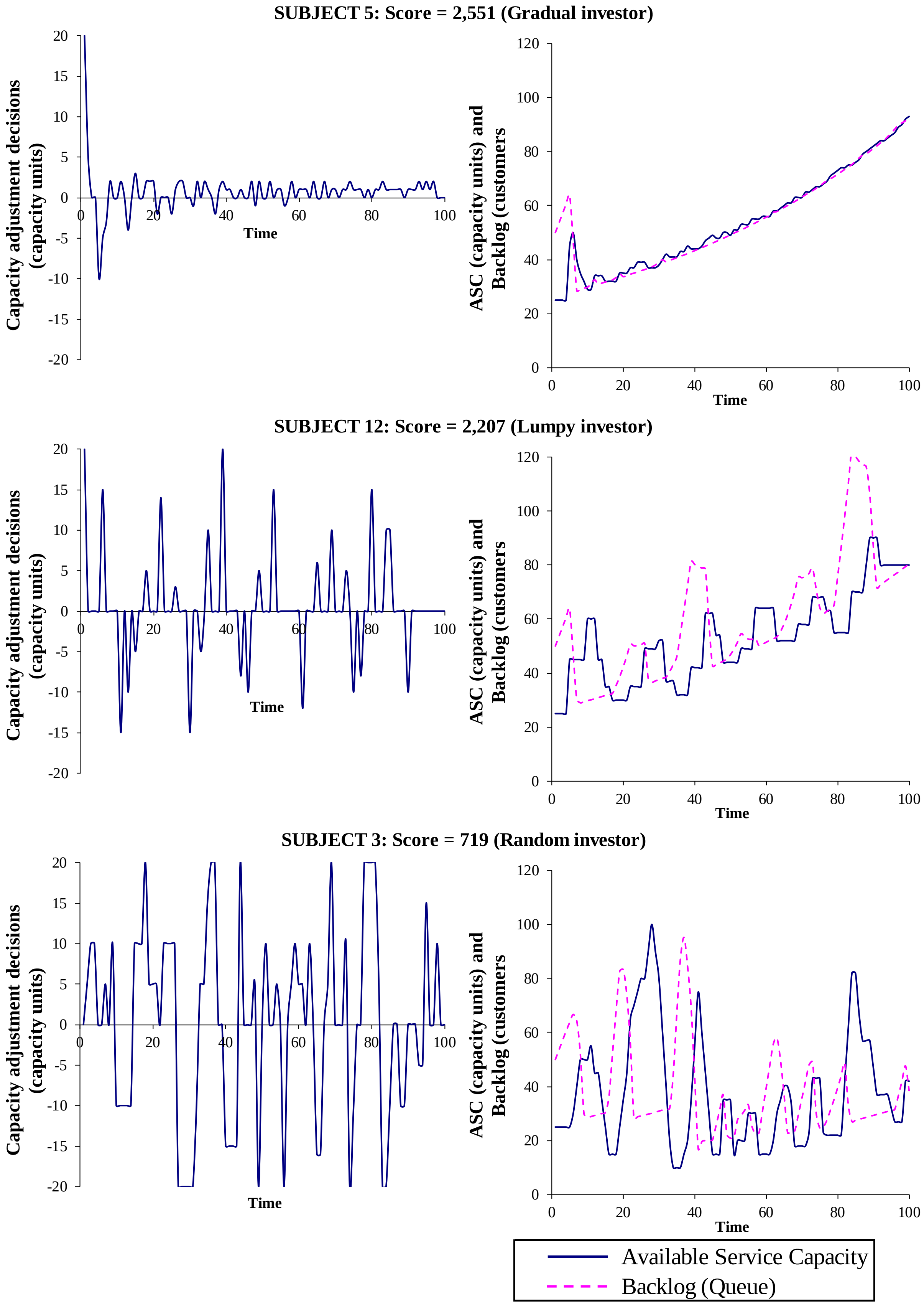


Figure 4.3 Typical behaviours observed in the lab.

Most subjects initially decide to increase capacity and later overreact to the decreasing backlog by removing capacity. We can interpret these first reactions as a learning process in which subjects are trying to adapt to the system behaviour during a transient period. After this period, the three groups of subjects behave differently. The first group is composed of those subjects who overreact to the initial increase of the backlog but quickly switch to making repeated small decisions to gradually adjust capacity over time (e.g. Subject 5). Most of these decisions concern capacity orders (i.e. positive capacity adjustment decisions). Consequently, the garage's available service capacity for this kind of managers increases exponentially over time. The backlog increases at the same rate as the available service capacity. Thus, the sojourn time remains stable and the garage attract more customers. We consider that these subjects learn quickly from the system behaviour. This group achieved the higher scores for the experiment. We call these subjects "gradual investors".

The second group (e.g. Subject 12) represents those subjects who, after an overreaction to the initial backlog, make fewer but more aggressive capacity adjustment decisions than the subjects of the first group. Moreover, they continue to overreact to the evolution of the backlog over time. Hence, these subjects alternately decide to increase and decrease capacity (the capacity adjustment decisions are positive and negative in Figure 4.3). These decisions result in an oscillating pattern for both the available service capacity and the backlog. The subjects in this group are called "lumpy investors". They decide to add chunks of capacity when the backlog increases. But at such time the average sojourn time is very high, which affects customers' perception. This causes some customers to leave. Hence when the capacity orders materialise (after the delays involved in the decision process) the backlog of customers is already decreasing and this new capacity speeds up this process. The opposite logic occurs when the backlog is low and the subjects decide to decrease capacity. Given that the capacity orders are generally higher than the retirement decisions, the available service capacity increases over time, but more slowly than for the first group.

The last group includes subjects who, even after the transient period, continue to overreact significantly in both directions to the evolution of the backlog (e.g. Subject 3). We label these subjects "random investors". Given this behaviour, we conclude that this group of subjects is unable to handle the delay structure inherent to the system. They performed poorly; they achieved the lowest payoffs and occasionally found themselves with zero service capacity before the end of the experiment. In the remainder of the chapter we will exclude from the analysis those subjects who reached zero service capacity before period 100.

Figure 4.4 shows the autocorrelation function (ACF) of the capacity adjustment decisions of some typical subjects. These graphs portray the autocorrelation of the subjects' decisions with the previous decisions at different time lags. For nineteen subjects out of 30, the first order lags are significantly greater than zero at the 5% level of significance, as shown in Figure 4.4 for subjects 3, 5, 11, and 30 (the first lag is outside the shaded region). This means that the decision which subjects make at time *t* is positively correlated with the one they made one period earlier (i.e. time *t-1*). This outcome suggests that the decisions taken one period earlier by subjects should be taken into account when estimating a heuristic to portray the way the subjects made their decisions. Note that the first order lags of subjects 12 and 18, who belong to the second group, are significantly different from zero, but they have other significant higher order lags. For instance, the eighth order lag of the ACF of subject 12 and the second, sixth, eighth and tenth order lags of subject 18. Similarly subjects 3, 5, 11 also present significant higher order lags. These significant higher order lags constitute an evidence of the presence of oscillating patterns in the subjects' decisions, such as shown in Figure 4.3. The lags within the shaded region are not significantly different from zero. For only four subjects out of 30 are all the lags non-significant. Three of these belong to the first group (gradual investors) and the fourth one to the second group (lumpy investors).

Next we propose a heuristic to model the way subjects made their decisions. This heuristic is formulated as a multivariate regression model which considers the existing autocorrelation in the capacity adjustment decisions of the subjects. First we estimate the parameters of this regression model. Next we incorporate this heuristic in the SD model and simulate it using the estimated parameters for each subject. We compare the simulated behaviour to the one observed in the laboratory.

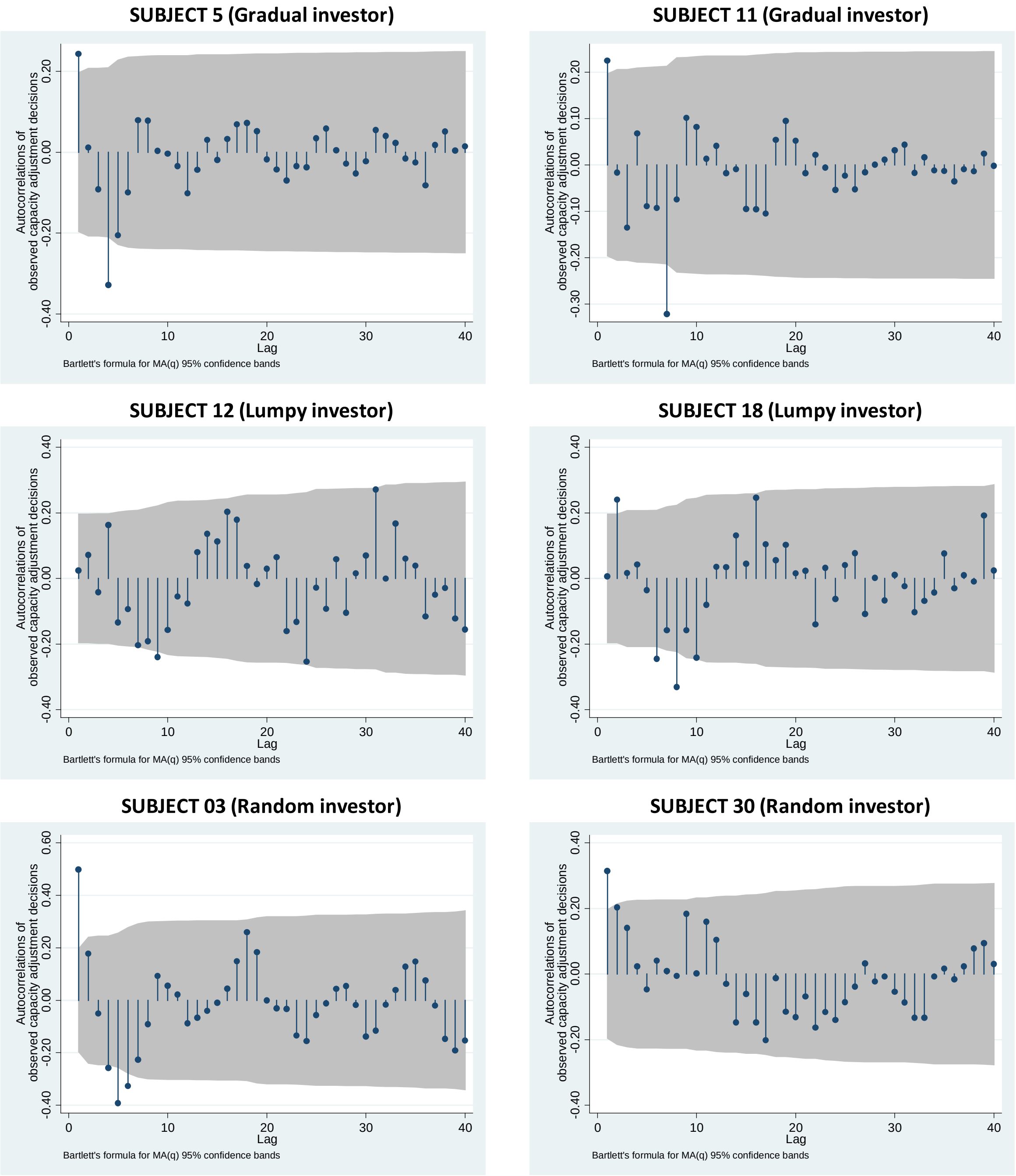


Figure 4.4 Correlograms of observed capacity decisions of some subjects.

4.4.1 Estimating a Decision Rule for the Capacity Adjustment

Following Delgado et al. (2011c), we assume that the manager attempts to adjust his capacity to a desired state, which is based on his estimate of the backlog. The manager determines his desired service capacity (DC_t) by computing the ratio between his perception of the backlog of work (EQ_t) and a market reference of the sojourn time (τ_{MR}). This market reference is assumed to be the same that current and potential customers use to base their decisions.

The perception of the backlog of work (EQ_t) can be estimated in different ways. In behavioural simulation there are several formulations to form expectations from recent observations. Sterman (1987, 1989) explains some of the most used formulations in SD. Adaptive expectations (Nerlove 1958), which is the most typical of these formulations, is the one van Ackere et al. (2010) use to model the manager's perception. Delgado et al. (2011c) test both adaptive and static expectations (Sterman 1989). Static expectations only consider the most recent information the manager has about the backlog of work (Q_t).

At the end of the experiment, we asked subjects to complete a questionnaire about the information they used most to take their decisions and how they used this information. 80% of the subjects considered the information about the backlog when taking their decisions. Thus, we hypothesise that the manager's desired service capacity is determined as a function of the backlog of work (Q_t),

$$DC_t = f(Q_t).$$

The goal of the manager is to maximise his profits and to do so he should satisfy his customers by avoiding high sojourn times. Thus we assume that the manager has his own reference sojourn time which does not necessarily equal the market reference. Then, a function for the desired service capacity of the manager could be:

$$DC_t = \delta * Q_t \tag{4.4}$$

where $1/\delta$ represents the reference average sojourn time of the manager. Once the manager knows his desired service capacity, he compares this value to what he considers to be his available service capacity in order to decide how much service capacity to order or to withdraw (recall that some of the managers' past capacity decisions are still to be implemented). When his desired capacity exceeds the current capacity, the manager decides to order capacity. When the opposite occurs he decides to remove capacity.

We formulate the hypothesis that subjects include their previous decisions, which have not been yet implemented, into their estimate of available service capacity. These not yet implemented capacity adjustments (ΔC_t) are given by the difference between the capacity on order (CO_t) and the capacity to be retired (CbR_t) at time t:

$$\Delta C_t = CO_t - CbR_t \tag{4.5}$$

And the future service capacity (FSC_t), once all past decisions (ΔC_t) are implemented, will be:

$$FSC_t = ASC_t + \Delta C_t \tag{4.6}$$

Previous research in capacity management suggests that decision makers often take into account only a fraction of the supply line when managing capacity adjustments (Sterman 1989a, 1989b. Hence, we incorporate the parameter ψ to represent the proportion of the not yet implemented previous decisions which the manager accounts for when taking his next decision. As in the previous chapter, we call this parameter the "coherence factor" and its interpretation is the same. So, the future service capacity (FSC_t), which the manager takes into account when taking capacity adjustment decisions, is modelled as follows:

$$FSC_t = ASC_t + \psi * \Delta C_t \tag{4.7}$$

Thus the perceived gap ($PGap_t$) between the desired service capacity and the future service capacity managers face each period is:

$$PGap_t = DC_t - FSC_t \tag{4.8}$$

Replacing DC_t and FSC_t using equations 4.4 and 4.7, we obtain:

$$PGap_t = \delta * Q_t - (ASC_t + \psi * \Delta C_t) \tag{4.9}$$

A manager might not wish to, or be able to, close this gap immediately as it is based on a perceived need. Thus, let χ be the speed at which the manager choose to close this perceived gap, i.e. how aggressive is he when making decisions. We define the capacity adjustment decision (CAD_t) as $CAD_t = \chi * PGap_t$. Then including χ in Equation 4.9, we obtain:

$$CAD_t = \chi * [\delta * Q_t - (ASC_t + \psi * \Delta C_t)] \tag{4.10}$$

where χ must be between 0 and 1. This adjustment involves either an increase in capacity (when $CAD_t > 0$), a decrease in capacity (when $CAD_t < 0$), or leaving capacity unchanged (when $CAD_t = 0$). Defining $\delta' = \chi * \delta$ and $\psi' = \chi * \psi$, we obtain:

$$CAD_t = \delta' * Q_t - \chi * ASC_t - \psi' * \Delta C_t \tag{4.11}$$

The parameters δ', χ and ψ' are estimated using a multivariate linear regression without intercept.

According to the theory ψ (i.e. the coherence factor) must be between 0 and 1 (Sterman 1989a). Then we expect ψ' to non-negative and $\psi' \leq \chi$. If $\psi' = \chi$, then $\psi = 1$ and the subjects account fully for the not-yet implemented decisions when taking new capacity adjustment decisions. If $\psi = 0$, the not-yet implemented decisions are ignored. Similarly, given that $1/\delta$ represents the average sojourn time the manager takes as reference to estimate his desired service capacity, we expect δ' to be non-negative.

Table 4.3 contains the parameter estimates (i.e. δ', χ and ψ') together with the p-values of the F-test for the significance of the linear regression for the 30 subjects of the base case. The F-test indicates whether there is a linear relationship between the desired capacity adjustment and the set of the variables considered, i.e. Q_t , ASC_t and ΔC_t (Montgomery and Runger 2003). Unlike the t-test, which assesses if the individual effects of these variables are significant, the F-statistic is used to test the null hypothesis that these effects are simultaneously equal to zero. The R^2 is not included in this analysis, given that we are considering a model without intercept. Hence, the analysis of such a measure is not appropriate (Greene 2002 and Chatterjee and Hadi 2006).

According to the p-values of the F-test, the hypothesis that all parameters of the model are simultaneously equal to zero can be rejected for all but 2 subjects. In most of the cases, this hypothesis is rejected at the 1% of significance level. This means that at least one of the variables Q_t , ASC_t and ΔC_t contributes significantly to the model (Montgomery and Runger 2003). That is, the capacity adjustment decisions can be estimated by a linear model in terms of Q_t , ASC_t and ΔC_t for 28 out of 30 subjects Only for two subjects (7 and 29), the estimated parameters are simultaneously not significantly different from zero. Thus, we should assume that a linear model, which depends on Q_t , ASC_t and ΔC_t, is not appropriate to estimate the decided capacity adjustments of these subjects.

Concerning the significance of the individual parameters, we can observe that the majority of the estimates for δ' and χ are significant. Only for subjects 1 and 20, are both parameters not significantly different from zero while for subjects 19 and 29 solely δ is not significant. A different situation occurs with ψ', which is only significant for a little less than half of all subjects. This means that less than half of the subjects consider their not-yet implemented decisions when deciding new capacity adjustments.

Subject	δ'	χ	ψ'	F-Test
1	-0.02	-0.03	0.07	0.017 b
2	0.24 a	0.22 a	-0.08	0.000 a
3	0.19 a	0.21 a	-0.06	0.000 a
4	0.22 a	0.20 a	0.09	0.000 a
5	0.37 a	0.35 a	0.31 a	0.000 a
6	0.13 a	0.13 a	-0.22 a	0.000 a
7	0.04 b	0.07 b	0.05	0.171
8	0.12 a	0.11 a	-0.01	0.000 a
9	0.05 b	0.04 a	-0.34 a	0.000 a
10	0.29 a	0.26 a	0.17 c	0.000 a
11	0.13 c	0.11 c	0.05	0.000 a
12	0.22 a	0.23 a	0.38 a	0.000 a
13	0.24 a	0.25 a	0.25 a	0.000 a
14	0.20 a	0.17 a	0.18 b	0.000 a
15	0.15 a	0.18 a	0.24 b	0.000 a
16	0.16 a	0.24 a	0.20 b	0.000 a
17	0.15 a	0.14 a	-0.13 c	0.000 a
18	0.24 a	0.23 a	0.35 a	0.000 a
19	0.02	0.04 a	-0.38 a	0.000 a
20	0.04	0.03	-0.14 b	0.012 b
21	0.11 a	0.10 b	0.12	0.000 a
22	0.04 b	0.04 b	0.11	0.087 c
23	0.22 b	0.20 b	0.09	0.003 a
24	0.12 a	0.11 a	-0.02	0.000 a
25	0.28 a	0.27 a	-0.01	0.000 a
26	0.20 a	0.18 a	0.04	0.000 a
27	0.16 b	0.14 b	0.13	0.036 b
28	0.11 b	0.09 b	0.02	0.000 a
29	0.06	0.06 c	-0.02	0.218
30	0.07 a	0.08 a	0.12	0.000 a

H_0 is rejected at a significance level of: [a] 1%, [b] 5% and [c] 10%

Table 4.3 Parameters estimates

In Table 4.4 we summarise the estimated parameters of Equation 4.10 (i.e. χ, δ and ψ) for the subjects whose parameter are all significant in Table 4.3. Recall that these parameters have an interpretation as long as they are non-negative and χ and ψ are less than or equal to 1.

We can observe that δ and χ satisfy this condition for all subjects in Table 4.4, while the conditions of ψ are only satisfied for three subjects (5, 10 and 16).

Subject	δ	χ	ψ
5	1.080	0.346	0.891
6	0.979	0.132	-1.640
9	1.300	0.038	-9.022
10	1.091	0.265	0.640
12	0.983	0.225	1.682
13	0.993	0.246	1.006
14	1.145	0.174	1.031
15	0.844	0.178	1.369
16	0.673	0.242	0.814
17	1.082	0.140	-0.896
18	1.017	0.235	1.494

Table 4.4 Significant estimated parameters of the decision rule (see Equation 4.10) for the subjects of the base case

Recall that 1/ δ is the sojourn time which the manager wishes to guarantee to his customers, i.e. it is the reference sojourn time the manager uses to estimate his desired service capacity (see Equation 4.4). Given that all values of δ in Table 4.4 are higher than 0.5, we can conclude that all subjects consider reference sojourn times lower than the market reference, which is equal to 2 months. The market reference is the sojourn time customers take into account when deciding whether to join or leave.

Considering χ, we can say that subjects are very prudent (i.e. slow decision makers) when making capacity adjustments based on their perceptions, since the values of this parameter are quite low, the highest being 0.35.

Finally, for the three subjects whose parameter ψ satisfies the theoretical conditions, we can say that according to the proposed decision rule they consider 89%, 64% and 81%, respectively, of their not-yet implemented decisions ($\varDelta C_t$) when taking a new capacity adjustment decision. Concerning the other subjects we cannot use this decision rule to make any inference about the way they consider $\varDelta C_t$ when adjusting capacity. Hence, in further work we propose to go study other decision rules which account for the way the subjects consider their not yet implemented decisions.

As a further test we have simulated the model using the decision rule with the estimated parameters for each subject. Figure 4.5 shows the comparison between simulated capacity adjustments and the ones decided in the laboratory by the three subjects analysed above as typical cases. The decision rule closely matches the decisions taken by subjects of the first group (e.g. subject 5). It is worth recalling that the subjects of this group achieved the better performance among the three groups. However, this decision rule does not perform very well for the other two groups of subjects as illustrated in Figure 4.5. Although the simulated capacity adjustments for subjects 12 and 3 exhibit oscillating patterns as observed in the laboratory, the amplitude and the period of these patterns are different, and for the third group the resulting capacity is also very different.

Although the F-test indicates that the proposed decision rule can be used to explain how most subjects take their decisions in the laboratory, this does not imply that this decision rule is the only one which can do so. Indeed, the comparison between the simulated and observed behaviours shows that this decision rule is not adequate to portray the way all subjects made their decision. Given the previous discussion about this decision rule, we consider that there is not enough evidence to conclude that the way subjects adjust their capacity can be modelled through a linear model in terms of the variables considered. Nevertheless, based on the autocorrelation analysis and the significance of δ and χ, we conclude that these variables have a certain influence on the decisions taken by the subject in the lab and must be considered in further work.

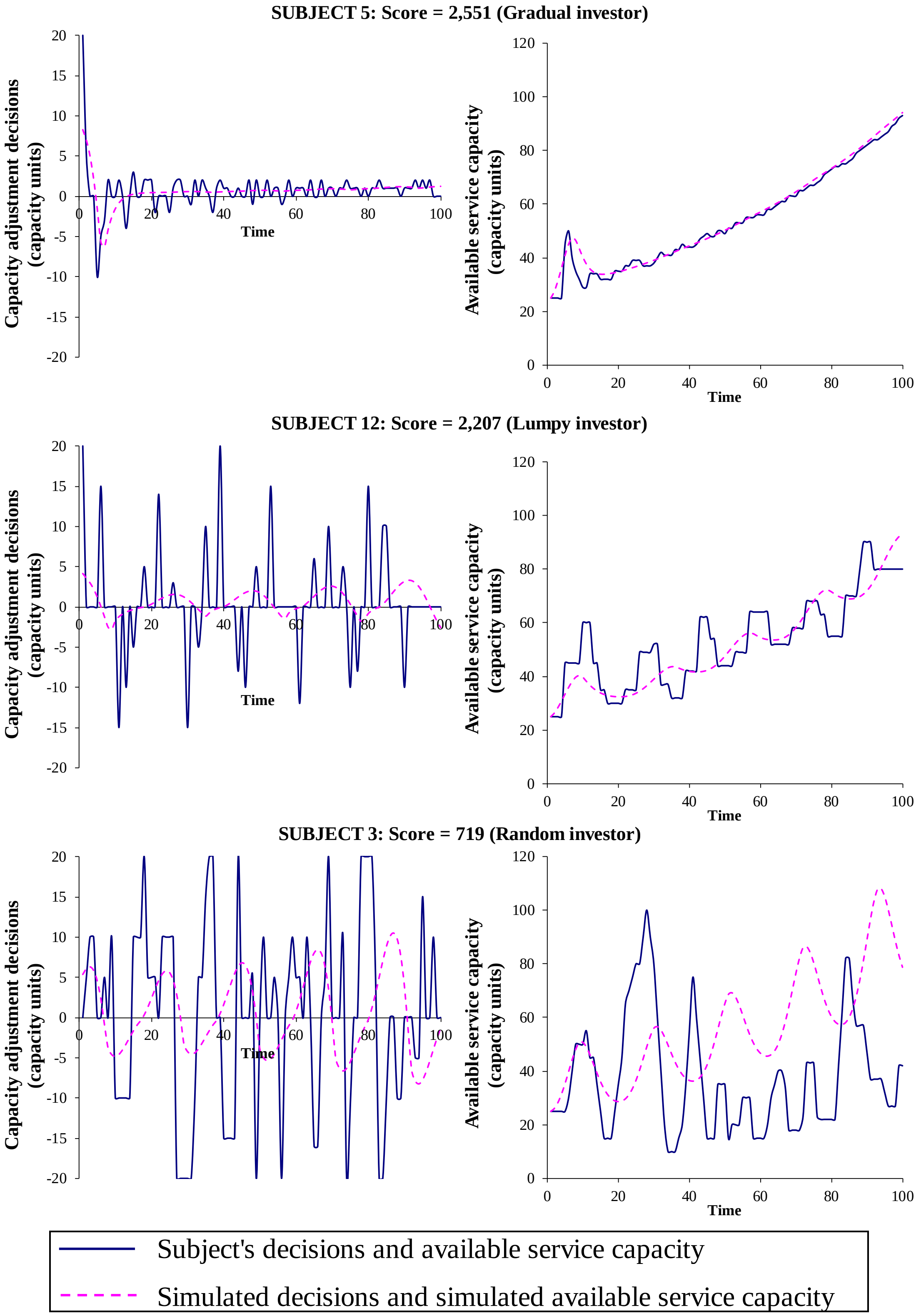


Figure 4.5 Comparison of experimental and estimated capacity decisions and available service capacity for the three typical subjects.

4.4.2 Treatment Results

Given that our data samples are very small, we should not use parametric tests to make inferences about the differences between the treatments. So we employ the nonparametric Mann-Whitney-Wilcoxon (MWW) test used in the previous chapter to compare the distributions of profits achieved in each treatment. Table 4.5 contains the corresponding p-values to test the null hypothesis that the median of the cumulative profits achieved by subjects assigned to the treatment in the row is equal to the median of the cumulative profits achieved by subjects assigned to the treatment in the column. Figure 4.6 shows the Box plots of the distributions of the cumulative profits achieved by the subjects for each treatment.

Treatment (X):	Treatment (Y):	Basecase	A	B	C
A	P-Value*	0.2805			
	P(X>Y)**	0.4190			
B	P-Value*	0.9035	0.1772		
	P(X>Y)**	0.4910	0.6020		
C	P-Value*	0.0008[a]	0.0029[a]	0.0000[a]	
	P(X>Y)**	0.2370	0.2680	0.1210	
D	P-Value*	0.0002[a]	0.0000[a]	0.0003[a]	0.0000[a]
	P(X>Y)**	0.7870	0.8280	0.7790	0.9320

* **H_0:** the median of the cumulative profits achieved by subjects assigned to the treatment in the row (X) is equal to the median of the cumulative profits achieved by subjects assigned to the treatment in the column (Y).

** Probability that the median of the cumulative profits achieved by subjects assigned to the treatment in the row (X) exceeds the median of the cumulative profits achieved by subjects assigned to the treatment in the column (Y)

[a] H_0 is rejected at a significance level of 1%

Table 4.5 Statistics of the Mann-Whitney-Wilcoxon (MWW) test for the comparison of the cumulative profits achieved in each treatment

Using the 1% significance level, we see clearly that the cumulative profits achieved in treatments C (i.e., high delays related to the decision process) and D (i.e., low delays related to the decision process) differ significantly from the ones achieved in the other treatments. Additionally we can observe that the subjects of treatment C typically achieve lower profits than the subjects of the other treatments. Conversely, the cumulative profits achieved in treatment D are generally higher.

The Box plots in Figure 4.6 illustrate the variability of the cumulative profits achieved by the subjects in each treatment and provide a comparison across treatments. The distribution of the cumulative profits of treatment D has the lowest variability, while the base case has the highest. Note also that despite some outliers, the cumulative profits achieved by subjects in treatment D are clustered around the median. The distributions of profits of treatments A, C, and D are more symmetric than those of treatment B and the base case. The distribution of the base case is the most asymmetric and it includes the poorest performance (see the lower outlier it is almost -$2,000). Nevertheless, Figure 4.6 confirms the conclusion of the MWW-tests that on average the poorer performances were achieved in treatment C and that the cumulative profits achieved in treatment D are on average the highest. So, we can conclude that the length of the delays involved in the implementation process of the decisions, which are known to the manager, strongly affect the cumulative profits which the garage can achieve. Moreover, the shorter these delays, the higher the cumulative profits of the garage and vice versa.

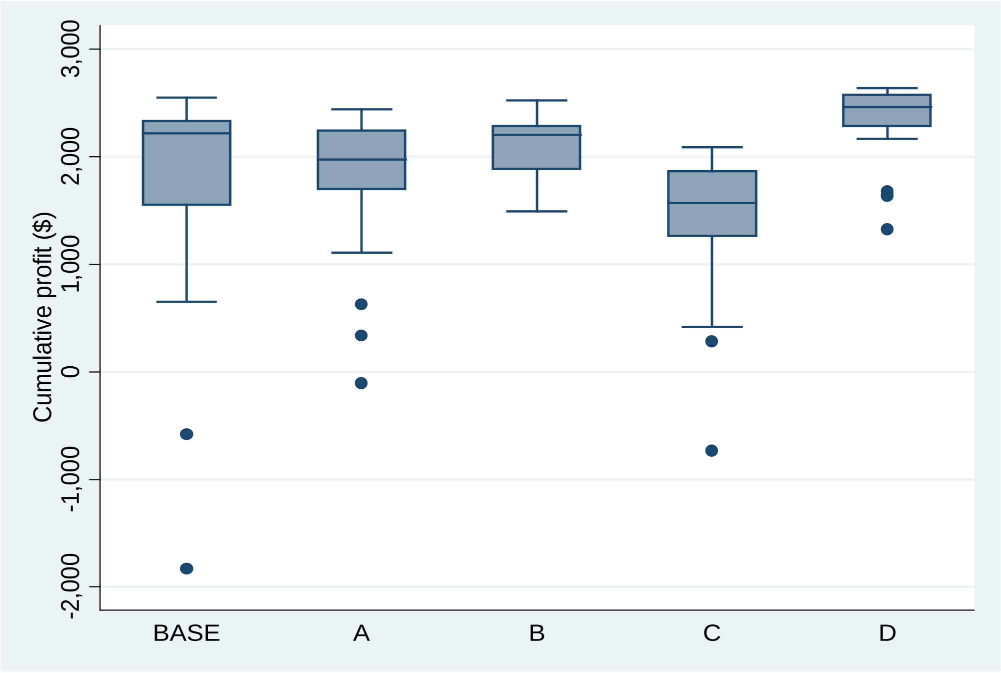


Figure 4.6 Box plots for the cumulative profits by treatment

4.5 Conclusions and Further Work

In this chapter, we have used a system dynamics (SD) model as an experimental platform to collect information about the way subjects taking the role of managers in a laboratory adjust the capacity of a service facility. The system was modelled as a garage for repair and maintenance of cars whose manager could adjust the capacity each period. The garage had a finite number of existing customers who patronised it and an infinite number of potential customers who each period could decide whether or not to join. Similarly, the existing customer could decide to leave. The decisions of both kinds of customers depended on their perceptions of the average sojourn time. Existing customers updated their perception based on their own experience, while potential customers did so through word of mouth. The goal of the subjects was to manage the service capacity of the garage in order to maximise their profits. To help subjects perform their task, they were provided each period with relevant information about the customer behaviour and the garage, such as the backlog of customers and the available service capacity.

Based on the results we can classify the subjects in three groups, whose decisions bring about similar behavioural patterns. In the first group are those subjects who gradually invest in service capacity over time. After initially overreacting to the increase of the backlog, these subjects learn from the system behaviour and continuously take small decisions to adjust the service capacity. This strategy enables them to achieve the better performances.

The second group represents those subjects who, after a slight overreaction to the initial increasing backlog, make fewer but more aggressive capacity adjustment decisions than the subjects of the first group. Their decisions exhibit an oscillating pattern: They alternate capacity orders and retirement decisions. Given that the former generally exceeds the latter, the available service capacity increases over time, but more slowly than that of the first group.

The last group includes subjects who, even after the transient period, overreact significantly to the backlog. This behaviour indicates that these subjects are unable to handle the delay structure inherent to the system. Hence, the garage of the subjects performs more poorly than that of the others and achieves the lower cumulative profits. Moreover, some of these subjects shut down their garage before the experiment ends.

The autocorrelation analysis shows that the capacity adjustment decisions made by most of the subjects are strongly positively correlated to the decisions taken one period early.

Given this analysis and the delay structure involved in the decision process, we developed a heuristic to model the decision rule applied by the subjects in the laboratory. This heuristic is a linear model which considers the current backlog, the available service capacity and the capacity decisions still to be implemented (once their delivery and dismantling delays expire). The significance tests of the estimated parameters indicate that when taking their current decisions only 13 out of 30 subjects in the base case account for their previous decisions, which have not yet been implemented. These tests also show that these subjects react prudently to changes in their perceptions when adjusting the service capacity. Additionally, when estimating their desired service capacity based on the backlog, these subjects target sojourn times which are below the market reference customers take into account for their decisions.

We have simulated the model using the estimated parameters and compared the simulated behaviour to the one observed in the lab. This decision rule fits very well for the subjects of the first group (those who gradually adjust capacity), but it does not capture the behaviour of the subjects of the other two groups. We thus propose as further work to delve more deeply into the statistical analysis of the information collected from the experiment in order to propose other decisions rules which capture the behaviour observed in the laboratory, and to extend this analysis to the other treatments. Another interesting issue will be to test the hypothesis that the autocorrelation of the subjects' decisions increases with the length of the delays.

Finally, the statistical comparison between different treatments, in which the length of the delays involved in the system was varied, enables us to conclude that the capacity implementation and dismantling delays have a significant impact on the cumulative profits achieved by the service facility. The longer these delays are, the higher the fraction of subjects who achieve low cumulative profits. In contrast, the different delays involved in the perception updating process of customers do not have a significant impact on the cumulative profits achieved by subjects.

Other extensions to this research include conducting further sets of experiments. For instance, incorporating a unit cost for each unit of capacity which the manager decides to add or remove; asking the manager for his expectations regarding the backlog; and including another group of human subjects who assume the role of customers.

5 SUMMARY OF KEY RESULTS AND SUGGESTIONS FOR FURTHER WORK

Studies of queueing phenomena have typically addressed the optimisation of performance measures (e.g. average waiting time, queue length and server utilisation rates) and the analysis of equilibrium solutions. The individual behaviour of the agents involved in queueing systems and their decision making process have been little discussed. Although this research has been useful for improving the efficiency of many queueing systems, or in other instances for designing new processes in social and physical systems, it has only provided us with a limited ability to explain the behaviour observed in many real queues.

In this dissertation we deviate from this traditional line of inquiry by analysing how the agents involved in the system make decisions instead of focusing on optimising performance measures or analysing equilibrium solutions. We have addressed three queueing problems dealing explicitly with the customers' and managers' decisions. These decisions are based on the expectations which customers and managers form regarding the state of the system. Customers and managers form these expectations using an adaptive expectation process whereby they update their memory using newly available information. They are considered conservative when they give more weight to their memory than to new information. In contrast, when they give more weight to new information, we say that they are reactive.

Traditionally, analytical modelling and simulation have been used to deal with queueing problems. The analytical approach describes mathematically the operating characteristics of the system in terms of performance measures, usually in "steady state". This approach is useful for low-complexity problems for which an analytical solution can be found with few simplifying assumptions. The second approach is more appropriate for complex problems, such as those where customers interact and share information regarding their experience in different queues and the decision process is subject to delays. A simulation approach enables modelling such problems in a more realistic way than the analytical approach, with fewer simplifying assumptions.

In chapters 2 and 3 we studied queueing systems which portray a captive market: customers periodically require a service and must choose which facility to join for this service. We have adapted the model proposed by Delgado et al. (2011a) (Appendix B) and Sankaranarayanan et al. (2011). We used a one-dimensional cellular automata (CA) to describe how customers interact with their neighbours and share information regarding their experiences. CA is an agent-based simulation methodology (North and Macal, 2007) in which agents are endowed with enough computational ability to update their state in the system by applying simple decision rules, such as joining the facility with the lowest expected sojourn time.

In Delgado et al. (2011b) (Appendix C), Delgado et al. (2011d) (Appendix E) and Chapter 2 of this thesis we incorporated uncertainty into the model proposed by Delgado et al. (2011a). Customers exhibit different degrees of risk-aversion which determine the extent to which they account for uncertainty when deciding which facility to patronise. The more risk-averse customers are, the more importance they give to uncertainty. Risk-neutral customers base their decision on the expected sojourn time (i.e. they ignore uncertainty), while risk-averse customers estimate an upper bound for the sojourn time using their expected sojourn time and their estimate of the level of uncertainty of this expectation. We studied the impact of the degree to which customers account for uncertainty when making decisions on the resulting collective behaviour and on the weighted average sojourn time of the system.

Interacting customers endowed with memory can yield different collective behaviours. Delgado et al. (2011a) discussed some of these behaviours and showed that the system achieves low weighted average sojourn times when groups of neighbours are loyal to a facility and customers are approximately equally distributed across all facilities. When an equal number of customers patronise each facility, and thus customers do not wish to change facility, the system achieves the Nash Equilibrium and this split yield the lowest weighted average sojourn time.

Delgado et al. (2011b) focused on the behaviour of risk-neutral customers and those with an intermediate degree of risk-aversion. They concluded that systems with customers having an intermediate degree of risk-aversion exhibit longer transient periods and converge more slowly to an almost-stable behaviour. Delgado et al. (2011d) showed that customers with a high level of risk-aversion experience, on average, low sojourn times when they are reluctant to incorporate new information when updating their expected sojourn time, whatever their attitude when updating their estimate of variance. Moreover, very risk-averse customers achieve their best performance when they are reactive to new information to update their perception of the variance. Customers with an intermediate level of risk-aversion experience low sojourn times when they are reluctant to update their expectations of both sojourn time and variance. Finally, risk-neutral customers and those with low risk-aversion achieve their best performance when they are most conservative with respect to updating their expected sojourn times.

In Chapter 2 we showed that there is a non-monotonic relationship between the degree of risk-aversion and system performance; customers with an intermediate degree of risk-aversion typically achieve higher sojourn time than the others and they rarely achieve the Nash equilibrium. Meanwhile, risk-neutral customers have the highest probability to achieve the Nash Equilibrium. Concerning the transient period, we extended the conclusion of Delgado et al. (2011b) to risk-averse customers in general, i.e. the more risk-averse the customers, the longer the transient period exhibited by the system. Indeed, risk-averse people take more volatile decisions. Consequently the system takes more time to achieve a stable behaviour.

While the optimal choice of updating parameters depends on the customers' risk attitude, service systems where customers are either close to risk-neutral or strongly risk-averse usually perform better than those who have an intermediate level of risk-aversion, whatever the parameter values.

In Chapter 3 we incorporated endogenous service rates into the model of Delgado et al. (2011a). Accordingly, we endowed the managers with the ability to adjust the service capacity. Each period they can decide to add or remove capacity based on their desired state and the current state of the facility. Each manager is characterised by a profile which depends on how conservative or reactive he is regarding new information, how rational when accounting for his previous decisions still in process of implementation and how fast he takes decisions. In that sense, we found that the more conservative a manager is regarding new information, the larger the market share his facility achieves. Additionally, the faster he takes decisions, the more likely he is to achieve a monopoly position. Another interesting result this model yields is the path-dependence phenomena of the managers' decisions. Once historical or random events determine a particular path, agents may become locked-in regardless of the advantages of the alternatives. In this sense, managers can have the same profile, but owing to the initial conditions of the model, their facilities can evolve completely differently. For instance, a system with three identical managers could become “locked-in” into a monopoly or duopoly situation.

The competition between managers causes the weighted average sojourn time of the system to converge to a benchmark value. This benchmark is a reference sojourn time managers use to estimate their desired capacity to satisfy the customers based on market expectations. In this sense, this benchmark is an attractor point of the system. The rate at which this convergence occurs depends on the managers and customers’ profiles. In some cases, the convergence period can be very long.

Chapter 4 tackled the third queueing situation we have addressed in this dissertation. This situation deals with the capacity management of a service facility with a finite number of current customers and an infinite number of potential customers. Customers are able to update their perceptions of the system based either on their experiences (current customers) or on information shared through word of mouth (potential customers). We focused on analysing the effects of the delays involved in the system. These delays are the time required by customers to update their perceptions and the implementation time of the manager’s decisions (i.e. delivery and dismantling delays). We have conducted a laboratory experiment in which human subjects take the role of a garage’s manager. In order to carry out this experiment we have applied the protocols of experimental economics (Friedman and Sunder 1994).

According to the results, we have classified the subjects in three groups, whose decisions bring about similar behavioural patterns and which we have labelled gradual investors, lumpy investors, and random investors. In the first group are those subjects who learn from the system behaviour and gradually invest in service capacity over time. This strategy enabled them to achieve the better performances. Subjects of the second group alternate their decisions between capacity orders and retirement decisions. They are more aggressive than the subjects of the first group. Given that the capacity orders usually exceed the retirement decisions, the available service capacity increases over time, but more slowly than that of the first group. In the last group are those subjects who are unable to handle the delay structure inherent to the system and take seemingly random decisions.

The experiment was performed for different treatments whereby we varied the length of the delays involved in the system. The statistical comparison tests across these treatments indicate that the delivery and dismantling delays significantly impact the cumulative profits achieved by the subjects in the laboratory. The longer these delays, the lower the cumulative profits typically achieved by the subjects. In contrast, the different delays involved in the updating process of the customers' perception did not significantly affect these profits.

Using autocorrelation analysis we have shown that the decisions taken by most of the subjects are positively correlated to the decisions taken one period early. We have also proposed a heuristic which models the subjects' decision rule based on their previous decisions still to be implemented, the current backlog and the available service capacity. This heuristic is formulated as a linear regression model without intercept. We have estimated the parameters of this decision rule and simulated the model using these estimates. We found that this decision rule fits very well for those subjects who gradually adjust capacity, but it does not capture the behaviour of the subjects of the other two groups. The significance tests of the parameter estimates indicate that when taking their current decisions only some subjects account for their previous decisions, which have not yet been implemented. Accordingly, we plan to delve more deeply into the statistical analysis of the data in further work in order to propose other decisions rules which capture the behaviour observed in the laboratory.

5.1 Overall Contribution

We hope that this research has provided some building blocks to incorporate behavioural aspects in queueing problems, thus contributing to the new field of behavioural operations management. We have built on and extended the work of van Ackere and Larsen (2004) and van Ackere et al. (2010). Using simulation models and experimental methodologies we have removed some simplifying assumptions of the classical queueing theory, such as steady-state condition, which enables us to be closer to reality. Particularly, the use of CA allows analysing the individual behaviour and the micro-dynamics of agents interacting in queueing systems. Throughout this thesis, we have considered that customers are autonomous and can decide each period which facility to join for service. This decision is based on the customers' most recent information, which they use to form expectations about the different facilities. Moreover, we have modelled the way customers account for the uncertainty involved in their perceptions and characterised customers through their risk-aversion level and the speed at which they update their perceptions. We showed that customers using adaptive expectations of waiting phenomena can generate different collective behavioural patterns. We have explained these patterns through the individual behaviour of customers.

Concerning queueing systems with endogenous service and arrival rates, this thesis shows that customers and managers' decisions exhibit a strong path-dependence phenomenon. Additionally we have shown that managers' decisions based on adaptive expectations cause the weighted average sojourn time of the system to converge to a market reference sojourn time.

The laboratory experiment of Chapter 4 enabled us to test hypotheses about the way human subjects, who play the role of managers of a service facility, make decisions regarding the capacity adjustment of this facility. These hypotheses have given us some insights, which theoretical models cannot confirm so far. For instance, the comparison of different treatments in Section 4.4.2 enables us to conclude that explicit delays for delivering and dismantling capacity in a service facility significantly affect the subjects' decisions, and thus the performance of the facility. Similarly we showed that the backlog, the available service capacity and the decisions, which have not yet been implemented, influence the way real subjects manage the capacity adjustment of service facilities.

5.2 Limitations

In this thesis we have restricted our work to a multichannel queueing system with three facilities and a neighbourhood size of 1. As mentioned before, considering a larger neighbourhood size with only three facilities would be to assume that customers could often have full information to take their decisions. We have performed some experiments simulating the one dimensional CA model with five and ten facilities. This structural change increases the system complexity drastically, since more facilities mean more states the cells can take. The results we found were intuitive: the greater the number of facilities the longer the transient period the system exhibits. Indeed, many facilities imply more experimentation and choice options. Accordingly, the learning process of customers in the system will take more time. Further work should consider higher dimensional CA models with a greater number of facilities (states) to analyse the impact of the available information and extend the theoretical framework for the analysis of collective behaviours and decision making in queueing systems.

The CA models we developed in this thesis assume that all customers arrive at the same time at the facility for service. This characteristic implies that our models are very specific and limits their application to situations occurring during rush hours, e.g. students who, every day during lunch time, must choose a restaurant for lunch.

A major limitation has been the computational time required to run many iterations. For instance, an iteration of the model used in Chapter 2 for 100 simulation periods takes around 3 seconds; the 10,201 iterations of the model required to obtain one panel of Figures 3.5 to 3.7 take around 50 minutes; the 9,000 iterations of the model required to compute all the estimates of Section 3.4.2.2 take 90 minutes. Extending this analysis to the 275,427 combinations considered in section 3.4.2.1 would require more than one year of simulations on a normal CPU. Most of our simulations were run using 24 CPUs simultaneously. An alternative solution could be to compile the code of the model using the Matlab compiler and run the model outside the MATLAB environment. This would reduce the computational time by running the model in shared libraries.

Although experimental economics is useful to collect information from a controlled economic environment and enables one to make conjectures about the subjects' behaviour, such conjectures only apply to the behaviour observed in the laboratory and cannot be extrapolated to reality. Indeed, "the relevance of experimental methods rests on the proposition that laboratory markets are "real" markets in the sense that principles of economics apply there as well as elsewhere" (Plott 1982, p 1520). The principle of experimental economics is to apply the simplicity of laboratory markets in order to reflect the way real people pursue real profits within a virtual context of real rules (Plott 1982). Within this context, the results of our experiment are only a first step towards understanding how managers adjust service capacity in situations where sojourn time is major factor of the customers' decision making process.

5.3 Further Work

The framework of behavioural operations management is still in its early stages and accordingly there is a large potential for further work that spans several research topics. Throughout this dissertation we have proposed many interesting alternatives to extend this work. Regarding the CA models, we have mentioned considering heterogeneous customers, i.e. customers with different degrees of risk-aversion and/or different levels of reactivity; modelling queueing systems where both customers and managers account for uncertainty; and assessing the influence of other service factors on the customers and managers' decisions, such as price and quality. Further extensions concerning the experimental field include incorporating unit costs for each unit of capacity which the manager decides to add or remove; asking the manager for his expectations regarding the backlog; and including another group of human subjects who take on the role of customers.

Other directions in which the work could be extended include introducing other typical characteristics of queueing systems which affect the customers' experience, such as balking, reneging and jockeying; analysing other decision rules; and studying other characteristics in the customers' and managers' profile.

One could also consider including a time dimension in the customers' choice, i.e. customers could choose the timeslot at which they would like to join the facility. In such a context the manager would have a limited scope to adjust the capacity of the different timeslots.

6 REFERENCES

[1] Agnew, C. E. “Dynamic modeling and control of congestion-prone systems.” *Operations Research* 24 (1976): 400-419.

[2] Albright, S. C. and Winston, W.L., *Management Science Modeling* (4rd ed.). Canada: South Western Cengage Learning, 2012

[3] Arango, S., J. A. Castañeda and Y. Olaya. “Laboratory experiments in the system dynamics field.” *System Dynamics Review* 28 (2012): 94-106.

[4] Arthur, W. B. “Competing technologies, increasing returns, and lock-in by historical events.” *The Economic Journal* 99 (1989): 116-131.

[5] Arthur, W. B. “Positive Feedbacks in the Economy.” *Scientific American* 262 (1990): 92-99.

[6] Bailey, N. “On queuing processes with bulk service.” *Journal of the Royal Statistical Society Series B* 16 (1954): 80-87.

[7] ———. “A continuous time treatment of a single queue using generating functions.” *Journal of the Royal Statistical Society Series B* 16 (1954b): 288-291

[8] Barrer, D. Y. “Queuing with impatient customers and indifferent clerks.” *Operations Research* 5 (1957a): 644-649

[9] ———. “Queuing with impatient customers and ordered service.” *Operations Research* 5 (1957b): 650-656.

[10] Bielen, F., and N. Demoulin. “Waiting time influence on the satisfaction-loyalty relationship in services.” *Managing Service Quality* 17 (2007): 174-193.

[11] Boots, N. K. and H. Tijms. “A Multi-server Queuing System with Impatient Customers.” *Management Science* 45 (1999): 444-448

[12] Borshchev, A. and A. Filippov. “From system dynamics and discrete event to practical agent based modelling: reasons, techniques, tools.” In *Proceedings of the 22nd International Conference of the System Dynamics Society, Oxford, England, July 2004.*

[13] Chamberlin E. H. “An experimental imperfect market.” *The Journal of Political Economy* 56 (1948): 95-108.

[14] Chatterjee, S. and A. S. Hadi. *Regression analysis by example* (4th ed.). New York: John Wiley & Sons, 2006

[15] Davis, M. M. and J. Heineke. “How disconfirmation, perception and actual waiting times impact customer satisfaction.” *International Journal of Service Industry Management* 9 (1998): 64-73.

[16] Delgado, C. A., A. van Ackere, E. R. Larsen, and K. Sankaranarayanan. “Collective behavioral patterns in a multichannel service facilities system: a cellular automata approach.” In *Operations Research, Computing, and Homeland Defense: Proceedings of the 12th INFORMS Computing Society Conference, Monterey CA, January 2011*. 16-27. Hanover, MD: INFORMS, 2011a

[17] ———. “Modelling decisions under uncertainty in a behavioural queuing system.” In *Proceedings of the 25th European Conference on Modelling and Simulation, Krakow, Poland, June 2011*. 34-40. Dudweiler, Germany: Digitaldruck Pirrot GmbH, 2011b.

[18] Delgado, C. A., A. van Ackere, E. R. Larsen and S. Arango. “Capacity adjustment in a service facility with reactive customers and delays: simulation and experimental analysis”. In *Proceedings of the 29th International Conference of the System Dynamics Society, Washington, D.C., July 2011*. Washington D.C.: The System Dynamics Society, 2011c.

[19] Delgado, C. A., A. van Ackere, and E. R. Larsen. “A queuing system with risk-averse customers: Sensitivity Analysis of Performance.” In *Industrial Engineering and Engineering Management (IEEM), 2011 IEEE International Conference on, Singapore December 2011*. 1720-1724. IEEExplore Digital Library, 2011d.

[20] Dewan, S., and H. Mendelson. “User delay costs and internal pricing for a service facility.” *Management Science* 36 (1990): 1502-1517.

[21] Edelson, N. M. “Congestion tolls under monopoly.” *The American Economic Review* 61 (1971): 873-882.

[22] Edelson, N. M. and D. K. Hilderbrand. “Congestion tolls for poisson queuing processes.” *Econometrica* 43 (1975): 81-92.

[23] Erlang, A. K. “The theory of probabilities and telephone conversations.” *Matematisk Tidsskrift* 20 (1909): 33-39.

[24] Forio Online Simulations. “Forio Simulate.” http://forio.com/simulate.

[25] Friedman, D., and S. Sunder. *Experimental methods: a primer for economists* (1st ed.). Cambridge: Cambridge University Press, 1994.

[26] Gallay, O. “Agent-based routing in queueing systems.” *PhD diss.*, Ecole Polytechnique Fédérale de Lausanne, 2010

[27] Gardner Jr., E. S. “Exponential smoothing: The state of the art—Part II.” *International Journal of Forecasting* 22 (2006): 637-666.

[28] Graves, S. C. “The Application of queuing theory to continuous perishable inventory systems.” *Management Science* 28 (1982): 400-406

[29] Greene, W. H. *Econometric analysis* (5th ed.). New Jersey: Prentice Hall, 2002

[30] Gross, D. and C. M. Harris. *Fundamentals of queueing theory* (3rd ed.). New York: John Wiley & Sons, 1998.

[31] Grossklags J. “Experimental economics and experimental computer science: a survey.” In *Proceedings of the 2007 workshop on Experimental computer science*, San Diego, CA, 2007.

[32] Ha, A. Y. “Incentive-compatible pricing for a service facility with joint production and congestion externalities.” *Management Science* 44 (1998): 1623-1636.

[33] Ha, A. Y. “Optimal pricing that coordinates queues with customer chosen service requirements.” *Management Science* 47(2001): 915-930.

[34] Hassin, R. and M. Haviv. *To queue or not to queue: Equilibrium behavior in queueing systems* (1st ed.) Boston: Kluwer, 2003.

[35] Haxholdt, C., E. R. Larsen and A. Ackere. “Mode locking and chaos in a deterministic queueing model with feedback.” *Management Science* 49 (2003): 816-830.

[36] Hui, M. K. and D. K. Tse. “What to tell consumers in waits of different lengths : an integrative model of service evaluation.” *The Journal of Marketing1* 60 (1996): 81-90.

[37] Ilachinski A. *Cellular Automata: a discrete universe* (1st ed.).Singapore: World Scientific Publishing Co. Pte. Ltd., 2001

[38] Kampmann, C. P. E. “Feedback complexity and market adjustment: an experimental approach.” *PhD diss*., Massachusetts Institute of Technology, 1992.

[39] Kendall, D. G. “Some problems in the theory of queues.” *Journal of the Royal Statistical Society. Series B* 13 (1951): 151-185.

[40] Koole, G. and A. Mandelbaum. “Queueing models of call centers : an introduction.” *Annals of Operations Research* 113 (2002): 41-59

[41] Koopman, B. O. “Air-terminal queues under time-dependent conditions.” *Operations Research* 20 (1972): 1089-1114

[42] Larson, R. C. “Perspectives on queues: social justice and the psychology of queueing.” *Operations Research* 35 (1987): 895-905.

[43] Law, A. K. Y., Y. V. Hui, and X. Zhao. “Modeling repurchase frequency and customer satisfaction for fast food outlets.” *International Journal of Quality & Reliability Management* 21 (2004): 545-563

[44] Leeman, W. A. “The Reduction of queues through the use of price.” *Operations Research* 12 (1964): 783-785.

[45] ———. “‘Comments’ on Saaty’s ‘The Burdens of queuing charges’.” *Operations Research* 13 (1965): 680-681.

[46] Liebowitz S. J. and S. E. Margolis. “Path dependence, lock-in, and history.” *Journal of Law, Economics, & Organization* 11 (1995): 205-226.

[47] Mendelson, H. “Pricing computer service: queuing effects.” *Communications of the ACM* 28 (1985): 312-321.

[48] Mendelson, H. and S. Whang. “Optimal incentive-compatible priority pricing for the M/M/1 Queue.” *Operations Research* 38 (1990): 870-883.

[49] Montgomery, D. C. and G. C. Runger. *Applied statistics and probability for engineers* (3rd ed.). New York: John Wiley & Sons, 2003.

[50] Moxnes, E. “Not only the tragedy of the commons: misperceptions of feedback and policies for sustainable development.” *System Dynamics Review* 16 (2000): 325-348.

[51] Naor, P. “The regulation of queue size by levying tolls.” *Econometrica* 37 (1969): 15-24.

[52] Nerlove, M. “Expectations and cobweb phenomena.” *The Quarterly Journal of Economics* 72 (1958): 227-240.

[53] Newbold, Paul. *Statistics for business and economics* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall, 1988

[54] North, M. J., and C. M. Macal. *Managing business complexity: discovering strategic solutions with agent-based modeling and simulation* (1st ed.). New York: Oxford University Press, 2007.

[55] Pullman, M. E. and G. M. Thompson. “Evaluating capacity- and demand- management decisions at sky resort.” *Cornell Hotel and Restaurant Administration Quarterly* 43 (2002): 25-36.

[56] Plott, C. “Industrial organization theory and experimental economics.” *Journal of Economic Literature* 20 (1982): 1485-1527.

[57] Rapoport, A., W. E. Stein, J. E. Parco and D. A. Seale. “Equilibrium play in single-server queues with endogenously determined arrival times.” *Journal of Economic Behavior & Organization* 55(2004): 67-91.

[58] Rapoport, A., W. E. Stein, V. Mak, R. Zwick and D. Seale. “Endogenous arrivals in batch queues with constant or variable capacity.” *Transportation Research Part B* 44 (2010): 1166-1185.

[59] Rump, C. M, and S. Jr. Stidham. “Stability and chaos in input pricing for a service facility with adaptive customer response to congestion.” *Management Science* 44 (1998): 246-261.

[60] Saaty, T. L. “The Burdens of queuing charges-comments on a letter by Leeman.” *Operations Research* 13 (1965): 679-680.

[61] Sankaranarayanan, K., C. A. Delgado-Alvarez, A. van Ackere, and E. R. Larsen. “The Micro-dynamics of queuing understanding the formation of queues.” *Working paper* (2011).

[62] Seale, D. A., J. E. Parco, W. E. Stein and A. Rapoport. “Joining a queue or staying out: effects of information structure and service time on arrival and staying out decisions.” *Experimental Economics* 8 (2005): 117-144.

[63] Senge, P. M. *The fifth discipline: The art and practice of the learning organization* (1st ed.). New York: Doubleday, 1990

[64] Sinha, S. K., N. Rangaraj and N. Hemachandra. "Pricing surplus server capacity for mean waiting time sensitive customers." *European Journal of Operational Research* 205(2010): 159-171

[65] Smith, V. L. "Microeconomic systems as an experimental science." *The American Economic Review* 72 (1982): 923-955.

[66] StataCorp.2011. (2011). Stata Statistical Software: Release 12. College Station, TX: StataCorp LP.

[67] Stein, W., A. Rapoport, D. Seale, H. Zhang and R. Zwick. "Batch queues with choice of arrivals: Equilibrium analysis and experimental study." *Games and Economic Behavior* 59 (2007): 345–363.

[68] Sterman, J. D. "Expectation formation in behavioral simulation models." *Behavioral Science* 32 (1987): 190-211.

[69] ———. "Modeling managerial behavior: misperceptions of feedback in a dynamic decision making experiment." *Management Science* 35 (1989a): 321-339.

[70] ———. "Misperception of feedback in dynamic decision making." *Organizational Behavior and Human Decision Processes* 43 (1989b):301-335.

[71] ———. *Business Dynamics: Systems thinking and modeling for a complex world* (1st ed.). Chicago, IL: Irwin-McGraw Hill, 2000.

[72] Stidham, S. Jr. "Optimal control of admission to a queueing system." *IEEE Transaction on Automatic Control* AC-30 (1985): 705-713.

[73] ———. "Pricing and capacity decisions for a service facility: stability and multiple local optima." *Management Science* 38 (1992): 1121-1139.

[74] Taylor, J. W. "Volatility forecasting with smooth transition exponential smoothing." *International Journal of Forecasting* 20 (2004): 273 - 286

[75] ———. "Invited comments on 'exponential smoothing: The state of the art - Part II' by E. S. Gardner Jr." *International Journal of Forecasting* 22 (2006): 671-672.

[76] Taylor, S. "Waiting for service : the relationship between delays and evaluations of service thanks." *The Journal of Marketing* 58 (1994): 56-69.

[77] The MathWorks Team. MATLAB 7.9 2009a. Cambridge, MA: The MathWorks, Inc, 2009

[78] Theil, H. and S. Wage. “Some observations on adaptive forecasting.” *Management Science* 10 (1964): 198-206.

[79] van Ackere, A. “Capacity management: pricing strategy, performance and the role of information.” *International Journal of Production Economics* 40 (1995): 89-100.

[80] van Ackere, A. and E. R Larsen. “Self-organising behaviour in the presence of negative externalities: A conceptual model of commuter choice.” *European Journal of Operational Research* 157 (2004): 501-513.

[81] van Ackere, A., C. Haxholdt and E. R Larsen. “Long-term and short-term customer reaction: a two-stage queueing approach.” *System Dynamics Review* 22 (2006): 349-369.

[82] van Ackere A. and Larsen E.R. “Queueing: a behavioural approach.” Swiss National Science Foundation (SNF) research proposal form No.100014_126584: 2009

[83] ———. “Dynamic capacity adjustments with reactive customers.” *Working paper*. University of Lausanne (2010).

[84] Wolfram, S. *Cellular automata and complexity: collected papers* (1st ed.). Boulder, CO: Westview Press, 1994.

[85] Yechiali, U. “On optimal balking rules and toll in the GI/M/1 queuing process.” *Operations Research* 19 (1971): 349-370.

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

7 APPENDIX

Appendix A. LIST OF PUBLICATIONS

1. Sankaranarayanan, K., E.R. Larsen, A. van Ackere and C.A. Delgado. “Genetic algorithm based optimization of an agent based queuing system”, In *International Conference on Industrial Engineering and Engineering Management (IEEM), IEEE International Conference on, Macao, China, December 2010*. 1344-1348. IEEEXplore® Digital Library, 2010

2. Sankaranarayanan, K., C.A. Delgado, A. van Ackere, and E.R. Larsen. “Behavioural queuing: an agent based modelling approach.” In *Proceedings of the IEEE International conference on Computer Modelling and Simulation (ICCMS 2011), Mumbai, India, January 2011*. 41-45. 2011 (Forthcoming in the *International Journal of Modelling and Optimization*)

3. Delgado, C. A., A. van Ackere, E. R. Larsen, and K. Sankaranarayanan. “Collective behavioral patterns in a multichannel service facilities system: a cellular automata approach.” In *Operations Research, Computing, and Homeland Defense. 12th INFORMS Computing Society Conference, ICS, Monterey CA, January 201*1. 16-27. Hanover, MD: INFORMS, 2011.

4. Delgado, C. A., A. van Ackere, E. R. Larsen, and K. Sankaranarayanan. “Modelling decisions under uncertainty in a behavioural queuing system.” In *Proceedings of the 25th European Conference on Modelling and Simulation, Krakow, Poland, June 2011*. 34-40. Dudweiler, Germany: Digitaldruck Pirrot GmbH, 2011.

5. Delgado, C. A., A. van Ackere, E. R. Larsen and S. Arango. “Capacity adjustment in a service facility with reactive customers and delays: simulation and experimental analysis.” In *Proceedings of the 29th International Conference of the System Dynamics Society, Washington DC. July 2011*. Washington, DC: The System Dynamics Society, 2011.

6. Delgado, C. A., A. van Ackere, and E. R. Larsen. “A queuing system with risk-averse customers: Sensitivity Analysis of Performance.” In *Industrial Engineering and Engineering Management (IEEM), 2011 IEEE International Conference on, Singapore, December 2011*. 1720-1724. IEEExplore Digital Library, 2011.

7. Sankaranarayanan, K., C. A. Delgado-Alvarez, A. van Ackere, and E. R. Larsen. “The Micro-dynamics of queuing understanding the formation of queues.” *Working paper* (2011).

8. Sankaranarayanan, K., C.A. Delgado, E.R. Larsen, and A. van Ackere. “Study on queuing behavior in a multichannel service facility using experimental methods.” In *Proceedings of the 1st International Conference on Operations Research and Enterprise Systems, ICORES 2012, Vilamoura, Portugal, February 2012.* 142-149. Portugal: SciTePress – Science and Technology Publications, 2012.

Appendix B.

“Collective behavioral patterns in a multichannel service facilities system: a cellular automata approach.”

Delgado, C. A., A. van Ackere, E. R. Larsen, and K. Sankaranarayanan,

In *Operations Research, Computing, and Homeland Defense. 12th INFORMS Computing Society Conference, ICS, Monterey CA, January 201*1. 16-27. Hanover, MD: INFORMS, 2011

© 2011 INFORMS | ISBN 978-0-9843378-1-1
DOI 10.1287/ ics.2011.0018

Collective Behavioral Patterns in a Multichannel Service Facilities System: A Cellular Automata Approach

Carlos A. Delgado A., A. van Ackere
HEC Lausanne, University of Lausanne, 1015 Dorigny – Lausanne, Switzerland
{carlos.delgado@unil.ch, ann.vanackere@unil.ch}

E.R. Larsen, K. Sankaranarayanan
Institute of Management, University of Lugano, 6904 Lugano, Switzerland
{erik.larsen@usi.ch, karthik.sankaranarayanan@usi.ch}

Abstract In this paper we propose a cellular automata model (CA) to understand and analyze how customers adapt their decisions based on local information regarding the behavior of the system and how the interactions of individuals and their decisions influences the formation of queues, which in turn impacts the sojourn time. We illustrate how a multichannel system of service facilities with endogenous arrival rate and exogenous service rate, based on local information and locally rational agents, may present different collective behaviors and in some cases reaches the Nash equilibrium.

Keywords Queuing; Simulation; Cellular Automata; Adaptive expectations, Collective behavior

1. Introduction

Queuing problems address a broad range of applications which have been widely tackled and discussed in various disciplines since Erlang (1909) [3], who is considered to be the father of queuing theory (Gross and Harris [5]), first published the telephone traffic problem. Studies of queuing systems encompass various disciplines including economics, physics, mathematics, and computer science.

Queuing is a fact of life that we witness daily and consider as an annoying situation. Banks, roads, post offices, and restaurants, are a few places where we experience queuing on a day-to-day basis. As the adage says, "time is money," is perhaps the best way of stating what queuing problems mean for customers. Queuing becomes an annoying and costly affair for customers who require a certain service routinely. In these cases, the experience enables customers to estimate the sojourn time for the next time, before deciding whether or not to join the queue and/or the best time to join, thus implying a dynamic queuing system with endogenous arrival rates which depend on the customers' expectations. For example, people who annually take their car to the garage for emissions tests, decide based on their experience what garage to take the car to and at what time to do so. Similarly, a worker or a student who daily has to select an hour and a restaurant to have lunch, has enough experience to choose the time and place that he considers less crowded.

The early works concerning queuing problems were confined to equilibrium theory (Kendall [9]) and focused on the design, running, and performance of facilities, with relatively little emphasis given to the decision processes of the agents of the system, i.e. the customers and managers of the facilities (van Ackere et al. [22]). Most queuing problems are tackled from an aggregated point of view. They are modeled by assuming static conditions, and exogenous arrival and service rates, and are analyzed in steady-state, despite the fact that they are dynamic and that the agents' decisions depend on the state of the system (Rapoport et al. [16]). More recently, researchers have attempted to shift the focus from these predominant assumptions of traditional queuing theory to a dynamic context in which agents' decisions are increasingly considered (e.g. Haxholdt et al [7] and van Ackere et al [21]). The present research is in this new direction.

There has been relatively little research aimed at analyzing and understanding the behavior of agents involved in a queuing system (van Ackere et al. [22]). The seminal papers on this subject are Naor [12] and Yechiali [24]. Koole and Mandelbaum [10] have suggested the incorporation of human factors as a challenge in order to advance the development of queuing models for call centers. Most of the models in this field are stochastic (e.g. Naor [12]; Yechiali, [23]; Dewan and Medelson [2]; Rump and Stidham [17]; Zohar et al [25]) and their form of feedback is either state-dependent (e.g. Naor [12]) or steady state (e.g. Dewan and Mendelson [2]). The stochastic models are aimed at understanding the impact of variability of the service and arrival processes on the system behavior (van Ackere et al. [22]). Some recent models are deterministic. For instance Haxholdt et al [7], van Ackere and Larsen [20] and van Ackere et al [21] analyze the feedback process involved in the customer's choice regarding which queue he should join in the next period. Haxholdt et al [7] and van Ackere et al [21] capture the average perceptions of the current customers in order to give feedback to the current and potential customers about the state of the system. van Ackere and Larsen [20] applied a single one-dimensional Cellular Automata (CA) model to capture the individual expectations of the customer about the congestion on a three road system.

We seek to understand how customers react to changing circumstances of the system. Our research involves studying the interactions of individuals within the system and the system's interactions with the individuals. These interactions are non-linear and involve feedback, and delays, and they reproduce adaptive and collective behaviors which depend on the initial values allocated to the customers. These issues make it difficult to solve models analytically; hence we adopt a simulation approach.

Specifically, we are interested in knowing how customers adapt their decisions based on local information regarding the behavior of the system. This information consists of their expectations (perceptions), their experiences, and that of their neighbors. In this way, we are moving the focus from analyzing the performance or designing the processes of a queuing system to analyzing the individual behavior of the agents and its impact on the system.

We apply agent-based simulation (North and Macal [15]), more precisely a CA model (Wolfram [23]), to capture the complexity of a self-organizing system. This complexity is represented by nonlinear interactions between the system's agents. "Cellular Automata are, fundamentally, the simplest mathematical representations of a much broader class of complex systems" (Ilachinski [8], p. 1). CA enables to endow agents with enough computational ability to interact with other agents of the system and share information. This is useful for modeling problems at any abstraction level (Borshchev and Filippov [1]). Taking into account the agents' autonomy, their interaction, and the fact that the information is shared between individuals at micro level, we consider that CA is a suitable methodology to help us model the system complexity. A CA model depicts agents interacting in a spatially and temporally discrete local neighborhood (Ilachinski [8]). The agents are represented as cells and each cell takes on one of k-different states at time t according to a decision rule (Ilachinski [8]). This decision rule determines the state of each cell at the next time period (t+1) based on the cell's current state and that of its neighbors (North and Macal [15]).

We use exponential smoothing (Gardner[4]) to estimate the agents' expectations of the congestion in the queues (in terms of sojourn time). In other words, the agents' decisions are based on adaptive expectations (Nerlove [14]). Exponential smoothing is based on a weighted average of two sources of evidence: one is the most recent observation and the other the estimation computed the period before (Theil and Wage [19]).

Consider a situation where customers routinely require a service and autonomously decide on a facility in a multichannel system with one queue for each channel (facility). There are also other applications in which customers do not choose a facility for service, but they may choose at what time to join the facility. In these cases we can consider each time period as a service channel. Once a customer is in the facility, if all servers are busy, customers must wait to be served. Their decision to return in the next period to the same facility, and therefore their loyalty, will depend on their past experience. Some examples of this kind of systems include an individual who must choose a garage for the inspection of his car, an individual who goes monthly to a bank to pay his bills, and an individual who goes to the supermarket weekly. In all these examples, the customer may choose the facility he wishes to be served at and at what time to do so. These are, in general terms, the kind of queuing problems to be studied in this research.

Simulating the CA model we found that it presents interesting collective behaviors of agents (customers) endowed with memory and local interactions with neighbors. In this paper we explain three of these behaviors: The first behavior depicts customers who switch between the different alternatives and do not achieve stability. The second behavior represents customers who alternate between two facilities, but the system achieves stability. In this case customers and their best performing neighbor alternate facility. The last behavior corresponds to a Nash equilibrium wherein after trying out several facilities, each agent remains loyal to one facility.

The paper is organized as follows. After this brief introduction, we provide a model description, which is followed by the simulation setup and results. We conclude the paper with comments and suggestions for future work.

2. The model

Consider a group of customers (referred to as agents) who routinely must choose which service facility to use in a multichannel system with one queue for each channel (facility). We assume an exogenous and identical service rate (μ) for all facilities, whereas the arrival rate (λ) is endogenous and depends on the agents' choice. They make their choice based on the sojourn time which they expect to face the next period at the different facilities. These expectations are built using the agents' most recent experience and that of their nearest neighbors. We apply a cellular automata model (CA) (Gutowitz [6], Wolfram [23]) to represent the interaction between agents and capture their expectations and dynamics. Agents are located in a one-dimensional neighborhood where each agent has exactly two neighbors, one on each side. The neighborhood represents, for instance, a social network encompassing colleagues, friends, people living next-door etc.

The structure of the model is assumed in the shape of a ring composed of cells. Each cell is an agent who may choose a service facility each time period. That is, the facilities are the states which each cell may take at each time period. Agents update their state through local interaction using a decision rule which is based on their own experience and that of their neighbors. In turn this experience depends on the state of all agents. We assume agents have a memory and the ability to update it using new information (previous experience). This memory contains the agents' expected sojourn time for the next period at the different facilities. We use adaptive expectations (Nerlove [14]) (also known as exponential forecasting (Theil and Wage [19]) or exponential smoothing (Gardner [4])) to model the updating process of agents' expectations. Such a CA model may be described as follows:

Let A be a set of n agents (cells) $\{A_1, A_2,\dots, A_i,\dots, A_n\}$ interacting with their neighbors and Q the set of m facilities (states) $\{Q_1, Q_2,\dots, Q_j,\dots, Q_m\}$ which agents (cells) may choose (take) at each time t. Agents interact in a neighborhood of size K (Lomi et al. [11]), which defines the number of neighbors on each side. For example, if $K=1$, agent A_i will interact with agents A_{i-1} and $A_{i+1.}$ Agent A_n will interact with A_{n-1} and A_1.

All m facilities have the same service rate μ, but different arrival rates (λ_j). Each agent A_i may join only one facility Q_j at each time t. We denote the state of agent A_i at time t by $s_i(t)$. Let S denote the set of states $s_i(t)$ of n agents at time t. This state $s_i(t)$ is one of the m possible facilities, that is, $\mathrm{S} \subset \{Q_1, Q_2,\dots, Q_j,\dots, Q_m\}$. Then the arrival rate (λ_{jt}) for the queue j at time t is a function of S, Q, and t. Let us consider the following function:

$$x_{ij}(t) = f(s_i, Q_j, t) = \begin{cases} 1 & \text{if } s_i(t) = Q_j \\ 0 & \text{otherwise} \end{cases} \qquad (1)$$

The arrival rate (λ_{jt}) for the queue *j* at time *t*, will be given by:

$$\lambda_{jt} = \sum_{i=1}^{n} x_{ij}(t) \qquad (2)$$

The state $s_i(t)$ for each agent A_i evolves over time according to the agents' expected sojourn time for each facility Q_j, denoted by M_{ijt}. At the end of each time period, the expected sojourn time of the agent for each facility is updated using two sources of information: his most recent experience and that of his neighbors A_{i-1} and A_{i+1} (Sankaranarayanan et al. [18]). Agent A_i's experience at facility Q_j at time *t* is denoted by W_{ijt}. Then, agent A_i's state ($s_i(t+1)$) and his expectation (M_{ijt+1}) for queue *j* for the next time period *t+1* are determined as follows:

$$s_i(t+1) = F(M_{ijt+1}) \qquad (3)$$

$$M_{ijt+1} = G(W_{i-K,jt}, \ldots, W_{ijt}, \ldots, W_{i+K,jt}, M_{ijt}) \qquad (4)$$

where $W_{i-K,jt}$ and $W_{i+K,jt}$ denote, respectively, the experience of neighbors A_{i-k} and A_{i+k}. The function G defines agent A_i's memory M_{ijt+1} (expectation) for queue Q_j for time *t+1*, using an adaptive expectations equation (Nerlove [14]), given by:

$$M_{ij,t+1} = \theta M_{ijt} + (1-\theta) W_{ijt}, \quad \theta \in (\alpha, \beta) \qquad (5)$$

where θ denotes the coefficient of expectations [14]. The parameter θ may take two different values depending on the source of information: When agents update their memory using their own experience, θ takes the value α. Otherwise, θ takes the value β. For $\theta = 0$, no weight is given to the past, which implies that the expected sojourn time equals the most recently experienced time. A value $\theta = 1$ implies no updating of expectations, i.e. the expectation will never change whatever the agent's new information. Thus, the higher the value of θ, the more conservative (or inert) the agent is towards new information, while a lower value means agents consider their recent information to be more relevant. The expected sojourn time for period *t*+1 (M_{ijt+1}) is thus an exponentially weighted average of the most recent experience W_{ijt} and the previous computed expectation (M_{ijt}). Agent A_i updates his memory in the following way:

(i) Based on his own experience (W_{ijt}), he will update his estimate of the sojourn time for his previously chosen service facility using $\theta = \alpha$.

(ii) The second source of information comes from the experience of the agent's neighbors $\{W_{i-K,jt}, \ldots, W_{i-1,jt}, W_{i+1,jt}, \ldots, W_{i+K,jt}\}$. He will update his memory for the previously service facility chosen by his best performing neighbor, i.e. the neighbor who has experienced the minimum sojourn time at the previous time period *W*, using $\theta = \beta$.

In the special case where the facility chosen by the agent and that chosen by his best performing neighbor coincide, the agent only updates his expectation once, using the minimum of α and β as weight. Regarding the decision rule, we consider rational agents who join the facility with the lowest expected sojourn time, that is, the agents update their state $s_i(t)$ each time period *t* using the minimum M_{ijt} according to equations (3) and (5). In special cases where an agent has the same expected sojourn time for two or more facilities and it is the lowest, he chooses as follows: if the expected time for the facility, which he chose, equals the minimum M_{ijt} he chooses this facility. If not, he checks whether the facility used by his fastest neighbor equals the minimum. If yes, he chooses this facility. Otherwise he chooses a facility at random: the facilities tied for the minimum expectation have equal probability of being selected.

Finally, we need to define the sojourn time W_{jt} at facility Q_j, given that λ_{jt} agents selected this facility at time t. Unfortunately, the steady state equations are only valid for queuing systems that reach equilibrium, and in which the average service rate exceeds the average arrival rate.

We need a congestion measure which can be used for our transient analysis where at peak times agents cluster in the same facility and the arrival rate temporarily exceeds the service rate. Considering the above, we use a congestion measure proposed by Sankaranarayanan et al [18] for a multichannel service facility with the same service rate (μ) for all facilities and endogenous arrival rate (λ_{jt}). Such a measure is given by:

$$W_{jt} = \frac{\lambda_{jt}}{\mu^2} + \frac{1}{\mu} \tag{6}$$

Then, by Little's law and the definition of ρ ($\rho = \lambda_{jt}/\mu$), the expected number of people for facility Q_j at time t is given by:

$$L_{jt} = \rho_{jt}(\rho_{jt} + 1) = \rho_{jt}^2 + \rho_{jt} \tag{7}$$

These measures satisfy the behavioral characteristics involved in the well-known Little’s Law and the steady state equations [5], but remain well-defined when ρ ≥ 1. For more details about the formulation of these measures, see Sankaranarayanan et al [18]. A brief description of the formulation and validation of Equations 6 and 7 is given in the Appendix A.

3. Simulation Setup

The agents of a CA model are endowed with memory (North and Macal [15]). This feature enables us to use this framework to investigate the problem we address here. We model the agents’ memory using adaptive expectations as described above. As the system behavior depends on the initial values of memory assigned to the agents, i.e. the evolution of the system is path dependent, our model cannot be solved analytically. Hence we use simulation to understand the system behavior. For its implementation we use Matlab, a numerical computing environment used in engineering and science.

The CA model is configured with 120 agents (i.e. the number of cells n in the one dimensional discrete lattice) and 3 facilities (i.e. number of states m which each cell may take). In this paper we use a neighborhood size (*K)* equal to 1, due to limited computational capacities. The service rate is the same for all facilities and equals 5 agents per unit of time. We simulate the model for 50 periods. These parameters are appropriate to observe the phenomena with which we are concerned. Each agent is allocated an initial memory for the expected sojourn time for each facility. These memories are distributed randomly around the optimal average sojourn time. In this paper we limit our study to the case where the agents use the same behavioral parameters value i.e. α = β = 0.5. All parameters used in this simulation are summarized in Table 1.

TABLE 1. Parameter values used for the simulation runs.

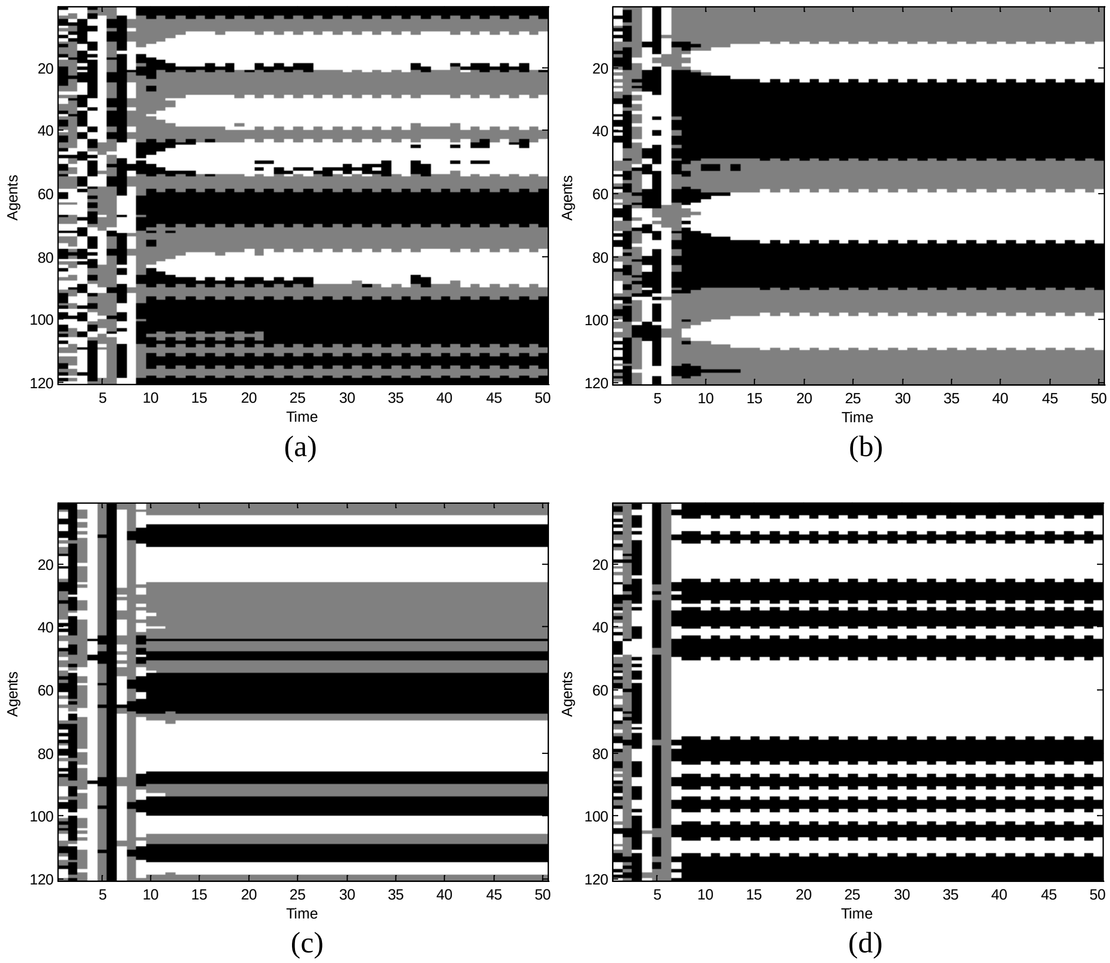
Parameter	**Description**	**Value**
m	Number of service facilities	3
n	Population size (Number of agents)	120
μ	Service rate	5
α = β	Weight to memory w.r.t. own experience and neighbors’ experience, respectively	0.5
Tsim	Simulation time	50
K	Neighborhood Size	1

4. Results

The four panels in Figures 1 to 3 illustrate different collective behaviors which may be captured by the CA model. We ran the simulation model using the same configuration for all runs, as shown in Table 1., but using different initial values of the expect sojourn times allocated to each agent. Recall that these values are assigned to each agent randomly.

We start by analyzing the more disaggregated results before studying the system globally. Figure 1 captures the evolution of the agents' choices of service facility over 50 time periods (one iteration) for 4 different initial values of expected sojourn times allocated to the agents. The horizontal axis represents time and the vertical axis the 120 agents. The colors indicate the state (chosen facility) of a particular agent at a particular time (black = facility 1, gray = facility 2 and white = facility 3).

FIGURE 1. Spatial-temporal behavioral evolution of agents' choice of service facility with $\alpha = 0.5$ and $\beta = 0.5$ with different values for the initial expected sojourn times.



(a)

(b)

(c)

(d)

We can observe that there is always an initial warm-up period whose length can vary. During this period, the agents try out the different facilities and all facilities are tested. We may say that agents are exploring the different facilities in order to learn from the system. For example in Figure 1b, agent 1 experienced the three facilities for the first five time periods with a sequence of 32231. This phenomenon depends strongly on the randomly allocated initial expected sojourn times. We can see that in many of these cases, some facilities are very crowded, implying that agents experience a large sojourn time at these facilities and expect the same situation for the next time period (e.g. Figures 1c and 1d show that facility 3 (white) is crowded at time 4). Consequently they move to another facility at the next period, generating the same problem for the new facility and in some cases forgetting the previous facility (e.g. in Figures 1c and 1d, no agents choose facility 3 at time 5, implying that one or both of the other facilities are crowded).

After the warm-up period, a set of more stable choices emerges over the next few periods. We can observe that agents present three different collective behaviors. The first is when there are still some agents moving through all facilities, as shown in Figure 1a. Figures 1b and 1d present the second case, in which a few agents keep switching between two facilities (e.g. in Figure 1d agents 98 and 102 switch between facilities 1(black) and 3(white) in a fairly regular pattern), while the others remain at the same facility. The logic behind this alternating behavior is that after the warm-up period, the sojourn times expected by a few agents at two facilities are very similar. As in this particular case agents give the same weight both to their own information and that of their neighbors, after updating their memories they consider that the facility which their neighbor used is more attractive than the one they patronize. They thus move to the neighbor's facility. A few agents moving to a facility during an almost stable period make it less attractive. Consequently, they decide to come back to their previous facility the next period, resulting in this switching behavior.

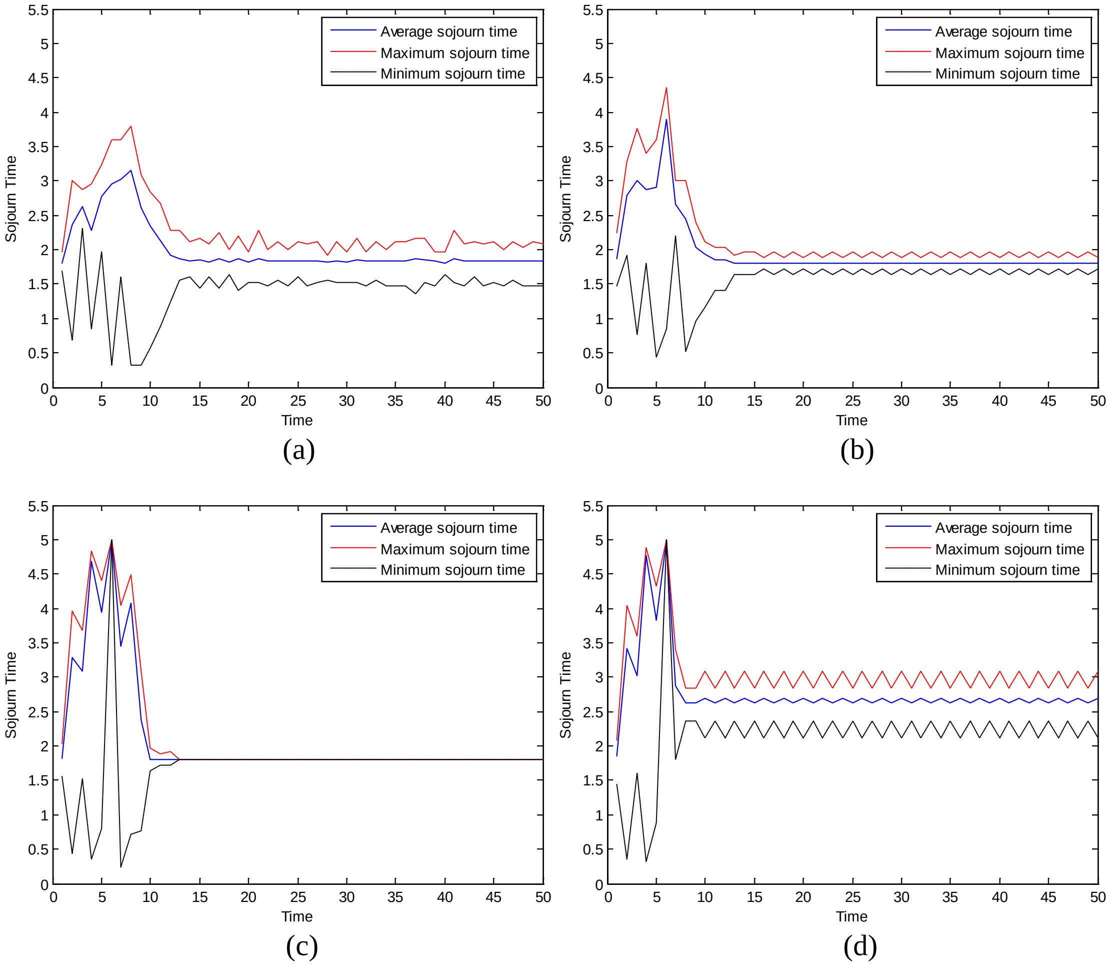
Figure 1d) also illustrates another phenomenon where one service facility is forgotten after the initial transition period. This particular case may occur when agents have had one or more very bad experiences at a facility. Its expected sojourn time becomes so large that none of the agents will patronize this facility for the next periods and they will thus be unable to update their expectation; hence the agents will never again use this facility in the future.

The final observed collective behavior is shown in Figure 1c: it portrays an equilibrium situation, which corresponds to the case where the agents are equally distributed across the three facilities (i.e. 40 agents at each facility), and all agents choose to remain at the same facility. They will stay at the same facility because once the system reaches steady-state they are in the facility which minimizes their expectation of sojourn time (i.e. maximization of their pay-off (Nash [13])) given the other agents' choices. That is, they reach the Nash equilibrium: each player's decision is optimal against that of the others (Nash [13]). Given that the three facilities are identical, an equal split of the agents across the three facilities is the only Nash equilibrium which the system can achieve. This situation coincides with the social optimum and yields a sojourn time of 1.8 time units. However, there are many ways in which agents can achieve this collective behavior (40 agents remaining at each facility over time). Which one materializes depends on the initial conditions. All facilities have the same sojourn time and the agents' estimates will converge to reality. Thus no agent wants to switch facility and this behavior will remain stable over time.

Figure 2 shows the evolution of the average sojourn time, along with the minimum and maximum sojourn times experienced by the agents for each time period. This figure provides us with a more aggregated view of the system's behavior. The major fluctuations occur during the warm-up period. For the four cases shown in Figure 2, the average sojourn time after the warm-up transition period are respectively 1.839, 1.802, 1.800, and 2.644. The first two are close to the Nash equilibrium (1.8), the third one confirms the equilibrium condition of the system and the latter is significantly higher than the Nash equilibrium. During the transition period, the average sojourn time of the system and the maximum sojourn time experienced by an agent are respectively 100% and 200% higher than the average sojourn time in steady state while the minimum sojourn time experienced by any agent is less than 50% of the average sojourn time in steady state.

The average sojourn time stabilizes after the transition phase. Note that once the system has stabilized, the average sojourn time of the system may oscillate as in Figures 2a and 2d. The same fluctuating pattern occurs with the maximum and minimum sojourn times in Figures 2a, 2b, and 2d. In general terms this fluctuating behavior occurs because a few agents keep changing facility, often alternating between two facilities, as illustrated in Figures 1a, 1b, and 1d. This behavior is not seen in Figure 2c because the system has reached the Nash equilibrium. While Figures 2b and 2d present a well-defined oscillating pattern, the oscillations in Figure 1a are irregular. This is because the expectations of some agents for the three queues are very similar; they thus keep facility, as shown in Figure 1a (e.g. agent 26 between times 36 and 38).

FIGURE 2. Four examples of average sojourn time for parameters $\alpha = 0.5$ and $\beta = 0.5$ with different values for the initial expected sojourn time



(a) (b) (c) (d)

Even though some agents in Figure 1b are switching between 2 facilities, the average sojourn time in steady state remains constant. This occurs because in one of the facilities (in this case facility 1) the number of agents stays constant and equals $\frac{n}{m}$, i.e. the number of agents in a facility when the Nash equilibrium is reached (40 for this case), while the other $n - \frac{n}{m}$ agents are divided among the other two facilities, with $\frac{n}{m} + v$ agents patronizing one facility, and the remaining $\frac{n}{m} - v$ the other one, where v is any integer number between 1 and n/m. For example in case (b) the number of agents in facility 2 alternates between 39 and 41 each time period. When 39 agents join facility 2, 41 join facility 3, and vice versa. In this case v equals 1.

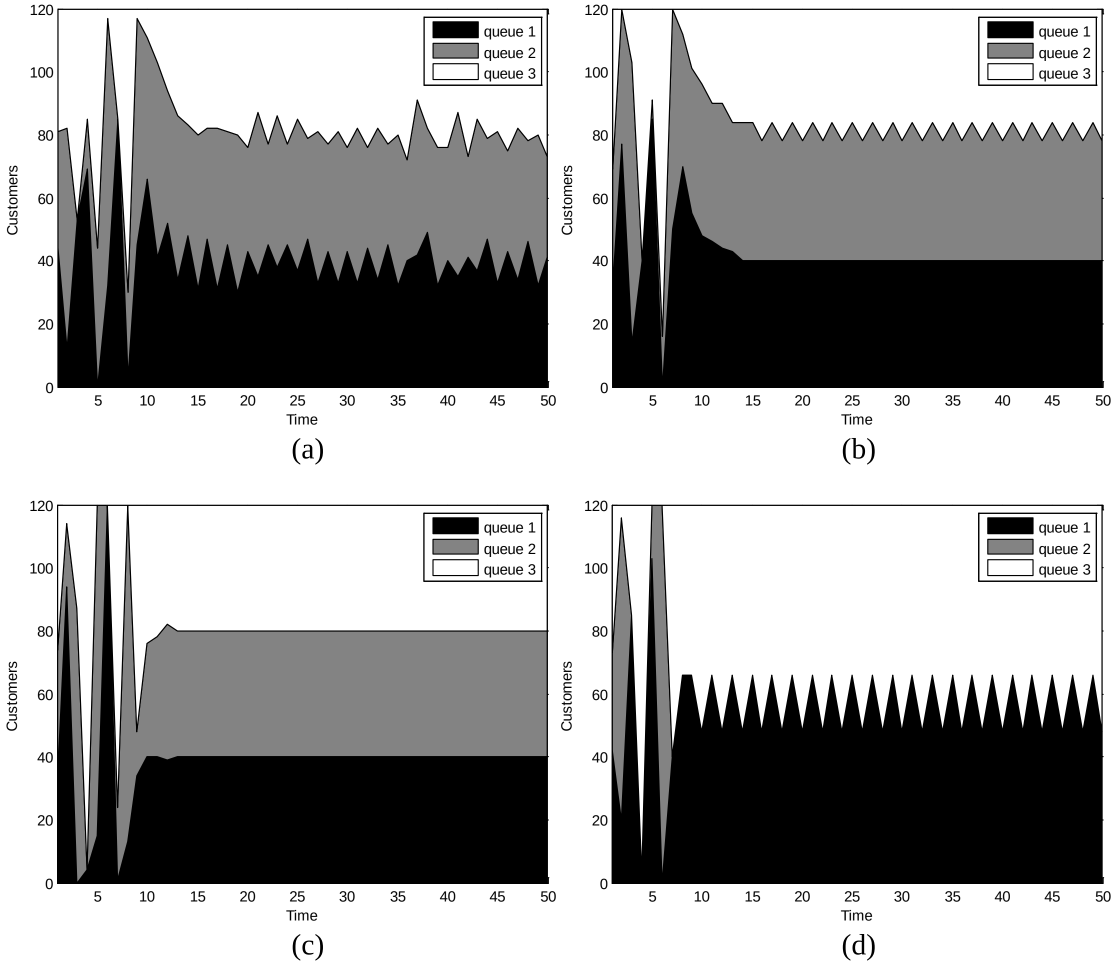
In Figure 2d there are just 2 facilities in use, facilities 1 and 3 (see Figures 1d and 3d). After the transition period the number of agents in each facility alternates each time period between $n_j + v$ and $n_j - v$ agents, n_j being the average number of agents who patronize facility j (i.e. j equals 1 and 3). When 66 agents join facility 1, the other 54 join facility 3, while when 48 agents go to facility 1, the other 72 join facility 3. Unlike Figure 2b, the average sojourn time in Figure 2d fluctuates because $n_1 \neq n_2$, i.e. the average number of agents (n_j) differs across facilities.

Figure 3 shows how agents are distributed across the different facilities over time. In these figures we analyze the system behavior at a macro-level. For instance, we easily can see when one facility is forgotten or which facilities are more crowded at a given moment of time, e.g. Figure 3c illustrates that facility 1 is forgotten at time 4, while at time 7 this is the only facility used by agents.

While Figure 2a indicates an almost stable average sojourn time, both Figures 1a and 3a confirm that there is no such stability at the micro level. In Figure 1a we saw how some agents switch between facilities. In Figure 3a we see that the distribution of agents in the three queues is changing over time in an irregular fashion.

Finally we can observe in Figure 3c that after the transition period the distribution of agents across the three queues does not change over time. This confirms that the Nash equilibrium may be reached with the parameter configuration used for this simulation, i.e. when agents give the same weight to both the memory and the new information (own experience and that of best performing neighbor).

FIGURE 3. Distribution of agents across the three service facilities for parameters $\alpha = 0.5$ and $\beta = 0.5$ depending on the initial expected sojourn time



(a) (b)

(c) (d)

5. Conclusions and Future Work

We have presented a one-dimensional cellular automata based queuing model to explain and understand how customers interact and make decisions in a multichannel service facility. We deviate from the traditional research approach to queuing which has mainly concentrated on the design, performance, and running of service facilities, assuming that customers' arrivals are exogenous and follow a stochastic process. We describe a self-organizing disaggregated queuing system with local interaction and locally rational agents (customers) who, based on their expectations (memory), decide which facility to join the next time period. They update their expectations based on two sources of information, their previous experience and that of their neighbors, using an adaptive expectation model.

Simulating this queuing model showed interesting collective behavior of agents (customers) endowed with memory and local interactions with neighbors. In this paper we have explained three of these. The first behavior depicts the case where customers do not find a facility that satisfies their requirements and continue to switch between alternatives. The second behavior represents the case where some customers have two preferred facilities and one of them corresponds to the one of their preferred neighbor (who has the better performance). In this case the customers alternate between 2 facilities. The last behavior corresponds to a Nash equilibrium wherein after trying out several facilities all agents find the most convenient one.

While the aggregated results (e.g. the evolution of average sojourn time) show that there is a certain stability in the system, the more disaggregated results (the agents' evolution in the system) may either contradict or confirm this analysis. By looking at the individual level we understand better how customers learn from the system and update their expectations regarding the system using the new information and their previously computed expectations (the memory). It also enables us to study how the customers' expectations may influence the stability of the system.

This is clearly a starting point for such a research agenda and we are working on extending the above mentioned framework. Extensions include playing around with different behavioral parameter values, considering service facilities of different sizes, including uncertainty in the customers' expectations, and also increasing the complexity of local interactions among agents i.e. changing the neighborhood parameter K. Another aspect would be to incorporate more decision capabilities into the model, such as decision making by services providers, i.e. considering that both the arrival and service rate are determined endogenously. An interesting approach would be to conduct experiments wherein human subjects act as customers so that we can verify the model and the heuristics that are used.

Appendix

A. Equation of the sojourn time (W_{jt}) (Adapted from Sankaranarayanan et al [18])

Let us consider an M/M/1 system (i.e. a one-server system with Poisson arrivals and exponential service times, see e.g. Gross and Harris [5]) in steady state. For such a system, the expected number of people in the system (*L)* satisfies equation (8):

$$L = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu - \lambda} \tag{8}$$

where ρ denotes the utilization rate λ/μ. Recalling Little's law

$$L = \lambda * W, \tag{9}$$

equations (8) and (9) imply that the average sojourn time in the system (*W)* equals

$$W = \frac{1}{\mu - \lambda}. \tag{10}$$

Unfortunately, these equations are only valid in steady state, which requires $\rho < 1$. We need a congestion measure which can be used for a transient analysis where at peak times the arrival rate temporarily exceeds the service rate. We have therefore attempted to identify a congestion measure that satisfies the behavioral characteristics of equations (8) to (10), but remains well-defined when $\rho \geq 1$. Such a measure should satisfy the following criteria:

(i) If ρ equals zero, the number of people in the facility, L, equals zero (Equation 8);

(ii) L increases more than proportionally in ρ (Equation 8);

(iii) When the arrival rate tends to zero, the sojourn time W is inversely proportional to the service rate μ (Equation 10);

(iv) When the arrival rate and service rate increase proportionally, leaving ρ unchanged, the waiting time W decreases (Equations 8 and 9);

(v) Little's Law is satisfied (Equation 9).

With these requirements in mind, we define L_{jt} as follows:

$$L_{jt} = \rho_{jt}(\rho_{jt} + 1) = \rho_{jt}^2 + \rho_{jt} \,. \tag{11}$$

Using Little's law and the definition of ρ yields the average sojourn time

$$W_{jt} = \frac{\lambda_{jt}}{\mu^2} + \frac{1}{\mu} \,. \tag{12}$$

Acknowledgments

This paper is based upon work supported by the Swiss National Science Foundation under grant No.100014_126584.

References

[1] A. Borshchev and A. Filippov. From system dynamics and discrete event to practical agent based modeling: reasons, techniques, tools. In *Proceedings of the Twenty-Second International Conference of the System Dynamics Society*, Oxford, UK. 2004. Wiley press.

[2] S. Dewan and H. Mendelson. User delay costs and internal pricing for a service facility. *Management Science*, 36(12): 1502–1517, 1990

[3] A.K. Erlang. The theory of probabilities and telephone conversations. *Matematisk Tidsskrift* 20(B): 33-39, 1909.

[4] E.S. Gardner Jr. Exponential smoothing: The state of the art – Part II. *International Journal of Forecasting* 22(4):637-666, 2006.

[5] D. Gross and C.M. Harris. *Fundamentals of Queueing Theory*. Wiley, New York. 3rd edition, 1998.

[6] H. Gutowitz. *Cellular Automata: Theory and Experiment*. The MIT Press, Boston, MA, 1991.

[7] C. Haxholdt, E.R. Larsen, and A. van Ackere. Mode locking and chaos in a deterministic queueing model with feedback. *Management Science* 49(6): 816-830, 2003

[8] A. Ilachinski. *Cellular Automata. A Discrete Universe.* World Scientific Publishing, Singapore, 1st edition, 2001

[9] D. Kendall. Some problems in the theory of queues. *Journal of the Royal Statistical Society Series B* 13(2):151-185, 1951.

[10] G. Koole and A. Mandelbaum. Queueing models of call centers: An introduction. *Annals of Operations Research,* 113(1): 41-59, 2002.

[11] A. Lomi, E R Larsen, A. van Ackere. Organization, evolution and performance in neighborhood-based systems. In Professor Brian Silverman, editor, *Geography and Strategy, Advances in Strategic Management,* Volume 20. Emerald Group Publishing Limited, Toronto, pp.239-265, 2003.

[12] P. Naor. On the regulation of queue size by levying tolls. *Econometrica* 36(1):15-24, 1969.

[13] J. Nash. *Non-cooperative games*. PhD, Princeton University, Department of mathematics, Princeton, NJ, 1950.

[14] M. Nerlove. Adaptive expectations and Cobweb phenomena. *The Quarterly Journal of*

Economics, 72(2):227-240, 1958.

[15] M.J. North and C.M. Macal. *Managing business complexity. Discovering strategic solutions with agent-based modeling and simulation*. Oxford University Press, New York, 1st edition, 2007.

[16] A. Rapoport, W.E. Stein, J.E. Parco, and D.A. Seale. Equilibrium play in single-server queues with endogenously determined arrival times. *Journal of Economic Behavior & Organization* 55(1): 67-91, 2004.

[17] C.M. Rump and S. Stidham Jr. Stability and chaos in input pricing for a service facility with adaptive customer response to congestion. *Management Science* 44(2): 246-261, 1998.

[18] K. Sankaranarayanan, C. A. Delgado, A. van Ackere, E. R. Larsen. The Micro-Dynamics of queuing understanding the formation of queues. Working paper, Institute of Management, University of Lugano, 2010.

[19] H. Theil and S. Wage. Some observations on adaptive forecasting. *Management Science* 10(2): 198-206, 1964

[20] A. van Ackere and E.R. Larsen. Self-organizing behavior in the presence of negative externalities: A conceptual model of commuter choice. *European Journal of Operational Research* 157(2):501-513, 2004

[21] A. van Ackere, C. Haxholdt, and E.R. Larsen. Long and short term customer reaction: a two-stage queueing approach. *System Dynamics Review*, 22(4): 349-369, 2006.

[22] A. van Ackere, C. Haxholdt, and E.R. Larsen. Dynamic capacity adjustments with reactive customers, Working paper 0814, Institute of Research in Management, Faculté des Hautes Etudes Commerciales (HEC). University of Lausanne, 2010.

[23] S. Wolfram. *Cellular automata and complexity*. Westview Press, Champaign, IL, 1st edition, 1994.

[24] U. Yechiali. On optimal balking rules and toll charges in the GI/M/1 queuing process. *Operations Research* 19(2): 349-370, 1969.

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

Appendix C.

“Modelling decisions under uncertainty in a behavioural queuing system.”

Delgado, C. A., A. van Ackere, E. R. Larsen, and K. Sankaranarayanan.

In *Proceedings of the 25th European Conference on Modelling and Simulation, Krakow, Poland, June 2011*. 34-40. Dudweiler, Germany: Digitaldruck Pirrot GmbH, 2011

MODELLING DECISIONS UNDER UNCERTAINTY IN A BEHAVIOURAL QUEUING SYSTEM

Carlos A. Delgado A. and Ann van Ackere
HEC, School of Business and Economics
University of Lausanne
Dorigny, 1015-Lausanne, Switzerland
E-mail: carlos.delgado@unil.ch,
ann.vanackere@unil.ch

Karthik Sankaranarayanan and Erik R. Larsen
Institute of Management
University of Lugano
6904, Lugano, Switzerland
E-mail: erik.larsen@usi.ch,
karthik.sankaranarayanan@usi.ch

KEYWORDS

Queuing problems, agent based simulation, cellular automata, adaptive expectations, uncertainty.

ABSTRACT

In this paper we use an agent-based modelling and simulation approach to model a queuing system with autonomous customers who routinely choose a facility for service. We propose a Cellular Automata model to represent the customers' interactions and study how customers use their own experience and that of their neighbours in order to update their memory and decide what facility to join the next period. We use exponential smoothing to update the customers' expected sojourn time. We incorporate uncertainty regarding these expectations into the customers' decision. We compare the resulting behaviour when customers take into account uncertainty to the case where they ignore uncertainty at both the individual and the system level.

INTRODUCTION

Queuing research has a wide gamut of applications spanning disciplines, which include telecommunications, informatics, economics, mathematics etc. Queuing theory is the mathematical analysis of queues (waiting lines) and the derivation of its performance measures (e.g. average waiting time). Erlang (Erlang 1909) the Danish engineer, who published his work on telephone traffic problem in 1909, is considered to be the father of queuing theory (Gross and Harris 1998).

Since then queuing has been studied extensively and tackled mainly with an aggregated view on the processes involved. However, queuing is a dynamic process where customers' decisions (disaggregated view) depend on the state of the system under study. Koole and Mandelbaum (2002) emphasize the need for including human factors in the context of call centre queuing models. Since Naor (1969) and Yechiali (1969) published their seminal papers, some researchers have tried to shift from the traditional approach, based on performance measures, to an approach wherein customers' decisions are incorporated into the queuing model.

Customers' decisions in queuing systems with stochastic arrival and service patterns have been studied theoretically in operations research and management science by Edelson and Hilderbrand (1975), Stidham (1985), Dewan and Medelson (1990), van Ackere et al. (1995), Rump and Stidham (1998), Zohar et al. (2002), among others. Hassin and Haviv (2003) discuss a vast literature, which depicts widely a framework on equilibrium behaviour in stochastic queuing systems. Comparatively, deterministic models have not been much discussed. Some models were proposed by Edelson (1971), Agnew (1976), Haxholdt et al. (2003), van Ackere and Larsen (2004), van Ackere et al. (2006).

Haxholdt et al. (2003) and van Ackere et al. (2006), using system dynamics, have included feedback into their model to look at the return rates of customers to the service facility. van Ackere and Larsen (2004) have used a one-dimensional cellular automata (CA) model in order to study formation of customers' expectations about congestion on a three-road system. Sankaranarayanan et al (2010a, 2010b) and Delgado et al. (2011) have applied this approach for a multichannel service facility with similar service rates for all facilities. They use an agent-based modelling (ABM) approach to understand and analyze how customers adapt their decisions based on adaptive expectations (Nerlove 1958). How local interactions among customers influence the formation of queues and how this formation is depicted by different collective behaviours, are explained. They use genetic algorithms to optimize the behavioural parameters of the agent-based model.

In this paper we propose an extension of the model of Sankaranarayanan et al. (2010a) and Delgado et al. (2011). We model a queuing system with endogenous and deterministic arrival rates using an agent-based simulation approach, more precisely a one-dimensional cellular automata, in order to study the collective behaviour of customers (referred to as agents in the remainder of the paper in order to contextualize the problem to the ABM methodology) who must choose routinely a facility for service. Some examples of this kind of systems include: a person who goes weekly to the supermarket, a person who must choose a garage for the inspection of her car, and a person who goes monthly to the bank to pay her bills.

We introduce uncertainty into the process of formation of agents' expectations in order to analyse how a risk-

Proceedings 25th European Conference on Modelling and
Simulation ©ECMS Tadeusz Burczynski, Joanna Kolodziej
Aleksander Byrski, Marco Carvalho (Editors)
ISBN: 978-0-9564944-2-9 / ISBN: 978-0-9564944-3-6 (CD)

averse attitude may affect collective behaviour. In this way we differ from Sankaranarayanan et al. (2010a) and Delgado et al. (2011) since the agents' decision policy in our model considers both the agents expectations and their uncertainty regarding those expectations. In order to model the agents' uncertainty we use the concept of volatility of forecast errors (e.g. Taylor 2004, 2006).

Our simulation results indicate that the transient period of the system is longer when its agents have an intermediate degree of risk aversion (R = 0.5), than when they are risk-neutral. After this transient period, if agents are risk-averse, the system converges more slowly to an almost-stable average sojourn time. Additionally, risk-agents are more likely to "forget" a facility, thus having a lower probability to be close to the Nash equilibrium (which requires using all facilities) when a steady state is reached. Systems where agents are either risk-neutral or strongly risk-averse perform better than those who have an intermediate level of risk aversion.

This paper is organized as follows: in the next section we describe the queuing system, the concepts of expectations and uncertainty, and how we use agent-based simulation and adaptive expectations to model the problem. In the third section we discuss and explain the simulation results. Conclusions and future work are presented in the last section.

MODEL DESCRIPTION

As pointed out earlier this paper builds on the work of Sankaranarayanan et al. (2010a) and Delgado et al. (2011); hence we briefly explain their model and then elaborate on how we incorporate the concept of uncertainty into the model.

Consider a fixed population of n customers (referred to as agents) who must routinely patronize a facility for service. Each period they face the decision of choosing among three facilities for this service. Each facility is set up with a queue. Queues are unobservable (Hassin and Haviv 2003), i.e. agents do not know the state of the system. Their decision therefore depends on their expectations about system congestion. The service rate (μ) is assumed to be fixed and identical for all facilities. The arrival rate is endogenous and depends on the agents' decisions. Reneging and balking are not allowed.

We use agent-based simulation to model this system; more precisely a cellular automata approach (Gutowitz 1991, Wolfram 1994) is adopted to model the interaction between agents, capture their expectations and analyse their collective behaviour. Agents interact in a one-dimensional neighbourhood assumed to have the shape of a ring, where one neighbour is located to the left and right of each agent. The last agent has to the first one and the last but one as neighbours. As we simulate a system with just three facilities, considering a neighbourhood size larger than 1 would be to assume that customers could often have full information. The neighbourhood can be assumed to represent, for instance, a social network encompassing colleagues, friends, people living next-door etc.

Each time period agents must choose a queue (state) following a decision rule, which is based on their expectations of sojourn time for the different queues and their uncertainty regarding these expectations. We assume smart agents who have the ability to remember information and update their memory using new information. This memory is updated using an adaptive expectations model (Nerlove 1958). Agents form expectations based on their last experience and that of their best performing neighbour, i.e. the neighbour who has experienced the minimum sojourn time in the previous period. Then, agents' expectations are adjusted using their uncertainty regarding such expectations. We model this uncertainty by applying the concept of volatility of forecasting errors (e.g. Taylor 2004). As variance is unobservable, exponential smoothing can be applied to estimate the squared residuals (Taylor 2006). Technically the model works in the following way.

Let A be a set of n agents (cells) $\{A_1, A_2,\ldots, A_i,\ldots, A_n\}$ interacting with their neighbours to select a facility among a set of m facilities (states) $\{Q_1, Q_2,\ldots, Q_j,\ldots, Q_m\}$ for service. The agents' decisions determine their state (queue) for the next period and this state will depend on their expected sojourn time for each facility, their uncertainty and how risk-averse they are. We denote expectations by M_{ijt} and uncertainty by s_{ijt}. R is a "risk aversion factor" which is identical for all agents. Let SM_{ijt} be an adjusted measure for the expected sojourn time of agent A_i at facility Q_j at time t. We call this measure "adjusted" because it incorporates the agents' uncertainty and their degree of risk aversion. The more risk-averse they are, the larger R is. SM_{ijt} is determined as follows:

$$SM_{ijt} = M_{ijt} + R{*}s_{ijt} \qquad (1)$$

SM_{ijt} can be interpreted as a form an upper bound of the agents' expected sojourn time. The agents' uncertainty, s_{ijt}, is assumed to be the volatility of their expectations and the risk aversion factor, R, may be considered as the how sensitive agents are to this volatility. In order to decide which facility to join, agents update this measure each period, based on the updated values of M_{ijt} and s_{ijt} as explained below.

Estimating the Expected Sojourn Times

The expectation M_{ijt} is updated using an exponential smoothing process, also called adaptive or exponential forecasting (e.g. Nerlove 1958; Theil and Wage 1964; Gardner Jr. 2006). It is a mathematical-statistical method of forecasting commonly applied to financial market and economic data, but it can be used with any discrete set of repeated measurements (Gardner Jr. 2006). This technique is based on weighted averages of two sources of evidence: The latest evidence (the most recent observation), and the value computed one period before (Theil and Wage, 1964). Exponential smoothing allows estimating the expected sojourn time, M_{ijt+1}, at time $t+1$ as a weighted average of the previous estimation of the expectation, M_{ijt}, and the recent observation W_{ijt} (experience). According to this, agents update their expectations for their chosen queue and that of their quickest neighbour using an exponentially weighted average with weight α. So, M_{ijt} can be expressed by:

$$M_{ijt+1} = \alpha M_{ijt} + (1 - \alpha) W_{ijt} \qquad (2)$$

where α is also known as the coefficient of expectations (Nerlove 1958) or smoothing parameter (Gardner Jr. 2006). See Delgado et al. 2011 for more technical details about the estimation of M_{ijt}.

Estimating the Uncertainty Measure for the Agents' Expected Sojourn Times

The measure of uncertainty, s_{ijt}, is modelled using the error of the estimation of M_{ijt}. According to Newbold (1988), if M_{ijt} is a smoothing estimation of W_{ijt}, the error in such an estimation will be:

$$e_{ijt} = W_{ijt} - M_{ijt-1} \qquad (3)$$

and the cumulative error on all expectations at time t could be estimated by the sum of the squared errors:

$$SS = \sum_{t=2}^{tsim} e_t^2 = \sum_{t=2}^{tsim} (W_{ijt} - M_{ijt-1})^2 \qquad (4)$$

where $tsim$ is the simulation time. However, in this way all the observations are given the same weight; a more realistic approach is to give a different weight to more recent errors. So in this context, the uncertainty, s_{ijt} , may be estimated using the concept of volatility forecasting which is calculated by means of the smoothing variance (e.g. Taylor 2006). As variance is unobservable, we can apply exponential smoothing in order to estimate this variance using the squared residuals (Taylor 2004). Thus, the smoothed variance, s_{ijt+1}^2, will be expressed as a weighted average of the previous estimate, s_{ijt}^2, and the new observation of the squared estimate error e_{ijt}^2. Thus, agents update the variance, s_{ijt}^2, as follows:

$$s_{ijt+1}^2 = \gamma * s_{ijt}^2 + (1-\gamma) * (W_{ijt} - M_{ijt-1})^2 \qquad (5)$$

where γ is the smoothing parameter (Taylor 2004). the volatility is then measured by the standard deviation s_{ijt}.

Once the uncertainty is estimated, agents consider this value in order to estimate an upper bound of their expected sojourn time for each facility. They then decide which facility to join the next period based on this adjusted estimate. Regarding the decision rule, we assume rational agents who always patronize the facility with the lowest upper bound for the expected sojourn time (SM_{ijt}), i.e. agents update their state by choosing the queue with the lowest value of SM_{ijt} in Equation (1). In the rare case where two or more queues have the same minimal adjusted expected sojourn time, agents choose among these facilities, giving first preference to their previously chosen queue and second choice to the one previously used by their best performing neighbour.

Average Sojourn Time in a Transient State

In this paper we consider a queuing system whose arrival rates may temporarily exceed the service rates. Hence we need a measure for the average sojourn time that enables us to study the system behaviour in a transient state, rather than in steady state

This measure is proposed and explained in Sankaranarayanan et al. (2010a). They consider a congestion measure which satisfies the well known Little's Law and the steady state equations (Gross and Harris 1998), but remains well-defined when $\rho \geq 1$ (Transient Analysis). Such a measure is given by:

$$W_{jt} = \frac{\lambda_{jt}}{\mu^2} + \frac{1}{\mu} \qquad (6)$$

where μ is the service rate for all facilities and λ_{jt} the number of agents arriving at Q_j at time t. For more details about the formulation of this measure, see Sankaranarayanan et al. (2010a) and Delgado et al. (2011). This measure is used throughout the paper.

SIMULATION RESULTS AND DISCUSSION

In order to study the impact of including uncertainty on the agents' decisions, we divide the simulation results in three parts: first we simulate a run of our chosen base case, which is described in Table 1. Using this simulation we compare the collective behaviour resulting from decisions, which include or ignore uncertainty. These results are reported in Figures 1 to 4. Then we study the distribution of average sojourn time for different values of the risk aversion factor, R, while keeping constant the smoothing parameters (Figure 5). Finally we perform a sensitivity analysis by considering combinations of smoothing parameters α and γ (Figure 6).

Table 1: Parameter values used for the simulation runs

Parameter	Value	Description
m	3	Number of service facilities
n	120	Population size
μ	5	Service rate
α	0.3	Weight to memory when updating the expected sojourn time
γ	0.7	Weight to memory when updating the estimated variance of the expected sojourn time
Tsim	100	Simulation time

Figures 1 and 2 exhibit the evolution of the agents' choices of service facility over 100 time periods for the same initial values of expected sojourn times allocated, which were allocated randomly to the agents. The horizontal axis depicts the time, while the 120 agents are represented on the vertical axis from top to bottom. The colours indicate the state (the chosen facility) of each agent each period. Black indicates facility 1, gray facility 2 and white facility 3.

Both figures exhibit an initial transient period. This period is longer when agents consider uncertainty in their decisions (around 34 periods, see Figure 2) compared to the case where uncertainty is ignored (around 15 periods, see Figure 1). This length can also vary depending on the randomly allocated initial expected sojourn times (Delgado et al. 2011). During this period agents explore all facilities in order to capture information about each one and try to learn from the system behaviour. The weight they give this information leads to herding: agents tend to imitate their best performing neighbours and at the end of the transient period one or two facilities tends to be crowded, as shown in Figure 1 (between periods 9 and 15) and Figure 2 (between periods 27 and 34). As more weight is given to new information than to memory (α is less than 0.5), agents react to the bad experiences by changing facility at the next period. For instance, in Figure 1 facility 1 is crowded at time 10, implying that no agents choose this facility in the next two periods. A similar situation occurs in Figure 2 at time 30 when all agents join facility 3 and it will not be used for the next three periods.

After the transient period a collective behaviour pattern starts to emerge. Small groups of agents tend to stay at the same facility over time, resulting in a certain degree of stability for the system. However the agents located on the borders between these groups keep shifting between two facilities (e.g. in Figure 1 agents 11 and 21 switch between facilities 1 (black) and 3 (white)). Between periods 15 and 65 agents on the borders change between facilities irregularly. After period 65 agents reach a regular pattern, which can be considered as stable. Comparatively, this behaviour cannot be identified in figure 2; at time 100, agents have not yet reached such a form of stability.

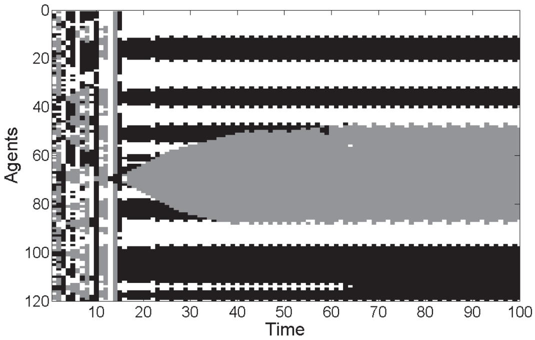


Figure 1. Spatial-temporal behavioural evolution of agents' choice of service facility with $\alpha = 0.3$; $\gamma = 0.7$ and R = 0

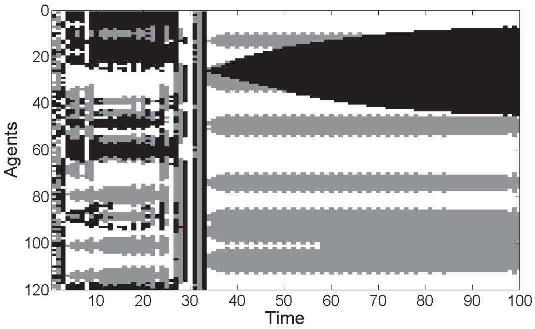


Figure 2. Spatial-temporal behavioural evolution of agents' choice of service facility with $\alpha = 0.3$; $\gamma = 0.7$ and R = 0.5

In both cases, two facilities have several small groups of loyal agents (facility 1 and 3 in Figure 1 and facility 2 and 3 in Figure 2), while one big group is loyal to the third facility (facility 2 in Figure 1 and facility 1 in figure 2). In Figure 1 this big group emerges at time 16 because when facility 2 reached its maximum level of congestion two periods earlier (at time 14), two agents did not use this facility. Their memory was therefore not affected by this bad performance, while that of the other agents increased a lot. At time 15 these two agents stayed at facility 1, which was then the most visited. Consequently their expectation for facility 1 soared. Thus at time 16, these two agents were the only ones to join facility 2, all other agents considering it a very poor choice. Hence, in this period of time they experienced an extremely low sojourn time as shown in Figure 3. They shared this experience with their neighbours over the next periods, who started coming back to facility 2. This increases the number of agents patronizing this facility, and thus the sojourn time and the agents expected sojourn time. At some point (around period 42) the last agents to join are disappointed and, given the information from their neighbours (who did not yet join), they leave again. These agents continue to alternate irregularly between this facility and that of their best performing neighbour until period 65. After this period the switching pattern stabilizes as discussed above. Agent 56 illustrates an odd case at time 64: facility 2 being close to its maximum occupation, the expected sojourn time of agent 56 for this facility increases and the minimum expected sojourn time of agent 56 at time 64 switches to facility 3 (in white). However, this agent experiences a bad sojourn time, which encourages him to come back to facility 2 and stay there for the remainder of the simulation.

A similar behaviour occurs in Figure 2, but the system requires more time to stabilize: no regular pattern is reached by the end of the simulation.

In Figures 3 and 4 we can observe that in those periods when most agents patronize the same facility the maximum, minimum and average sojourn time experienced by the agents reach extreme values. For instance, when agents ignore uncertainty (Figures 1 and 3) most agents choose facility 3 at time 13, two agents facility 1 and no agents facility 2. Hence, the two agents at facility 1 experience the lowest sojourn time (min $(W_{j,t}) = W_{1,14} = 0.28$), while the agents at facility 3 experience a very high sojourn time ($W_{3,14} = 4.92$). A similar case occurs in Figures 2 and 4, at time 34, a single agent joins facility 1 and experience the lowest sojourn time (min $(W_{j,t}) = W_{1,34} = 0.24$).

Figures 3 and 4 also illustrate that a system in which agents are risk-neutral converges faster to an almost-stable average sojourn time than a system with risk-averse agents.

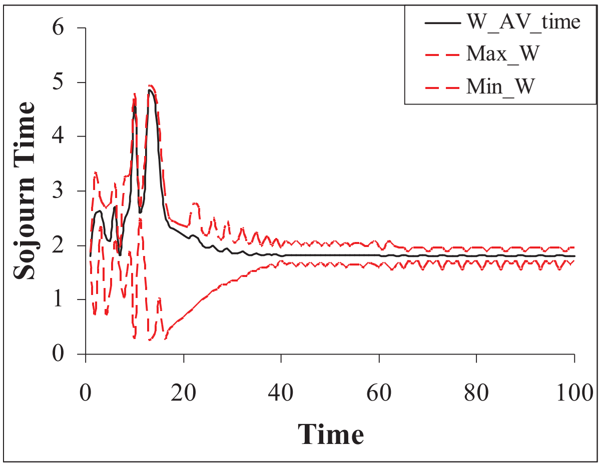


Figure 3. Average, maximum and minimum sojourn time for parameters $\alpha = 0.3$; $\gamma = 0.7$ and R = 0 (Without uncertainty)

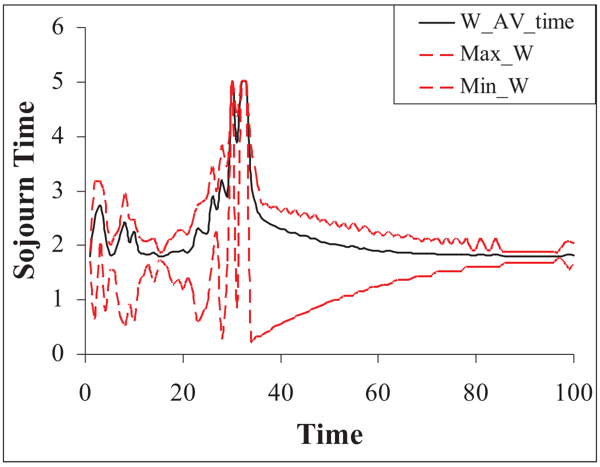


Figure 4. Average, maximum and minimum sojourn time for parameters $\alpha = 0.3$; $\gamma = 0.7$ and R = 0.5

Figure 5 shows the distributions of the average sojourn time for 1000 simulations of a system configured according to Table 1 with R equals to 0 (uncertainty is ignored) and 0.5 (a certain risk-averse attitude is considered). Each simulation was run over 1500 time periods and different initial expected sojourn time allocated to the agents randomly. The average sojourn time was calculated based on the last 500 periods of each run.

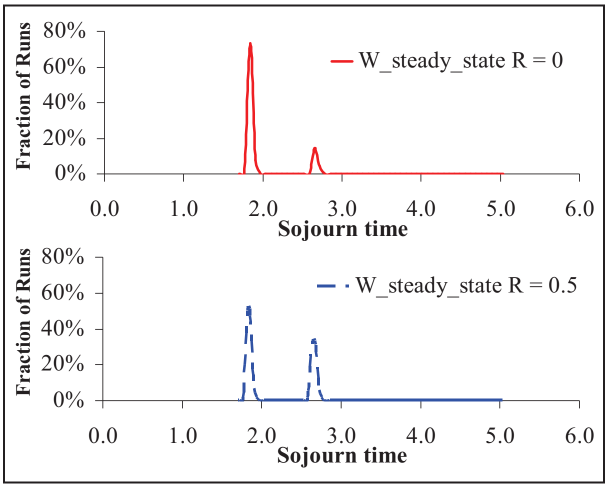


Figure 5. Distributions of average sojourn time for 1000 simulation with different initial conditions for α = 0.3; γ = 0.7

Figure 5 illustrates the dependence of the sojourn time on the initial allocated expected sojourn times and enables us to identify the probability that a queue will be ignored by the agents once the system reaches a steady state. The continuous line depicts the distribution of average sojourn times for the system where agents ignore uncertainty and the dashed one the system in which agents incorporate uncertainty into their decisions. The first peak of both distributions represents the proportion of cases where the three facilities are used in steady state. When R = 0, around 81% of the cases reach an average sojourn time in steady-state between 1.80 and 1.90 time units, compared to 57% of the cases for R = 0.5. The remaining cases (19% for R = 0 and 43% for R = 0.5) fall in the interval [2.58, 2.8]. In these cases one facility is ignored when the system has reached a steady state, i.e. all the agents are clustered in only 2 of the 3 facilities.

Figure 6 exhibits a sensitivity analysis. We consider several combinations of smoothing parameters α and γ, while varying the risk aversion factor from 0 to 1.5 in increments of 0.1. We simulate 1000 iterations for each combination of parameters (α, γ, R). Each iteration corresponds to a different set of initial expected sojourn times allocated to the agents randomly. All parameter combinations use the same random seeds in order to avoid a biased comparison.

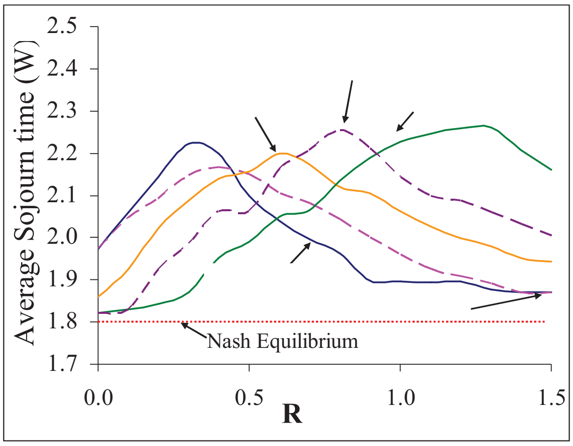


Figure 6. The Average Sojourn time in steady state as a function of R for selected values of α and γ. Each value represents the average of 1000 simulations.

Given that the three facilities have the same service capacity, the Nash equilibrium of the system is only achieved when agents are split equally across the three facilities. Such an equilibrium yields a sojourn time of 1.8 time units. Figures 6 indicates that on average the agents perform above the Nash equilibrium.

Figure 6 illustrates that when agents ignore uncertainty, that is R = 0, they perform better if they give at least the same weight to their memories as to the new information, i.e. if $\alpha \geq 0.5$. On the contrary, if $\alpha < 0.5$, very risk-averse agents (i.e. R is highest) perform better. That is, risk-neutral agents who are more conservative regarding new information (their experience and their neighbour's experience) achieve lower soujourn times on average than agents who give more weight to recent information. Likewise, agents attaching more importance to new information and having higher risk-aversion levels achieve sojourn times similar to those of conservative risk-neutral agents.

Next consider the parameter, γ, used in the variance smoothing process. Risk-averse agents must assign a high value to this parameter, i.e. $\gamma \geq 0.5$, if they want to perform well; otherwise they must use values α less than 0.5.

If agents use the same value for both smoothing parameters, those who are very risk-averse reach a better performance when they give more weight to new information, i.e. α and γ are both low.

In general terms, agents with an intermediate risk-aversion tend to perform more poorly than those who are either very risk-averse or close to being risk-neutral.

CONCLUSIONS AND FUTURE WORK

This paper uses a CA model to study the behavioural aspects involved in a queuing system in which customers decide what facility to choose for service each time period. They are endowed with memory and characterized by their risk attitude. They base their decisions on their expectation of the sojourn times and their uncertainty regarding these expectations.

Risk-neutral agents base their decision on the expected sojourn time, while risk-averse agents estimate an upper bound for the different sojourn times. When agents have an intermediate degree of risk aversion (R = 0.5), the system exhibits a longer transient period and, after this transient period, the system converges more slowly to an almost-stable average sojourn time. Additionally, risk-averse agents are more likely to "forget" a facility, thus having a lower probability to be close to the Nash equilibrium (which requires using all facilities) when a steady state is reached. Systems where agents are either close to risk-neutral or strongly risk-averse perform better than those who have an intermediate level of risk aversion.

The optimal choice of updating parameters depends on the agents' risk attitude. Consequently, future research will focus on studying the impact of the different behavioral parameters (α, γ, R). This will include giving different weights to own experience and information received from neighbors in the memory updating processes (expected sojourn times and variance). The next step will be to assume heterogeneous agents, i.e. agents with different smoothing parameters and risk factors in the same system. Finally, we plan to study the interaction between the decisions of managers and customers by allowing for facilities with different, adjustable service capacities.

REFERENCES

Agnew, C. E., 1976. Dynamic modelling and control of congestion-prone systems. *Operations Research* 24, No. 3, 400–419.

Delgado, C.A.; van Ackere, C.A.; Larsen, E.R.; and K. Sankaranarayanan. 2011. “Collective Behavioral Patterns in a Multichannel Service Facilities System: a Cellular Automata Approach”. In *Proceeding of the 12th INFORMS Computing Society Conference, ICS 2011* (Monterey, CA. Jan. 9- 11). INFORMS, Hanover, MD, 16-28.

Dewan, S. and Mendelson, H., 1990. User delay costs and internal pricing for a service facility. *Management Science* 36, No. 12, 1502–1517.

Edelson, N. M.,1971. Congestion Tolls under Monopoly. *American Economic Review* 61, No. 5, 873-882.

Edelson, N. M. and Hildebrand, D. K., 1975. Congestion Tolls for Poisson Queuing Processes. *Econometrica* 43, No. 1, 81-92.

Erlang, A.K. 1909. The theory of probabilities and telephone conversations. *Matematisk Tidsskrift* 20(B), 33-39.

Gardner Jr., E.S. 2006. Exponential smoothing: The state of the art – Part II. *International Journal of Forecasting* 22, No. 4(Oct-Dec), 637-666.

Gross, D. and Harris, C.M. 1998. *Fundamentals of Queueing Theory*. Wiley, New York. 3rd edition.

Gutowitz, H. 1991. *Cellular Automata: Theory and Experiment*. North-Holland, MIT Press, Boston, MA.

Hassin, R. and Haviv, M. 2003. *To Queue or Not to Queue: Equilibrium behavior in Queueing systems*. Kluwer, Boston, MA.

Haxholdt, C.; Larsen, E.R.; and A. van Ackere. 2003. Mode locking and chaos in a deterministic queueing model with feedback. *Management Science* 49, No. 6, 816-830.

Koole, G. and Mandelbaum, A. 2002. Queueing models of call centers: An introduction. *Annals of Operations Research* 113, No. 1, 41-59.

Naor, P. 1969. On the regulation of queue size by levying tolls. *Econometrica* 36, No. 1, 15-24.

Nerlove M., 1958. Adaptive expectations and Cobweb phenomena. *The Quarterly Journal of Economics* 72, No. 2, 227-240.

Newbold, P. 1988. *Statistics for Business and Economics.* Prentice Hall, Englewood Cliffs, N.J. 2nd ed.

Rump, C. M. and Stidham, S. Jr. (1998). Stability and chaos in input pricing for a service facility with adaptive customer response to congestion. *Management Science* 44, No. 2, 246-261.

Sankaranarayanan, K.; Delgado, C.A.; van Ackere, A.; and E. R. Larsen. 2010a. “The Micro-Dynamics of queuing understanding the formation of queues”. *Working paper*, Institute of Management, University of Lugano.

Sankaranarayanan, K.; Larsen, E.R.; van Ackere, A.; and Delgado, C.A. 2010b. “Genetic Algorithm based Optimization of an Agent Based Queuing System". In *Proceedings of the IEEE Industrial Engineering and Engineering Management Conference 2010*, (Dec. 7-10 Macau). IEEE*Xplore*® Digital Library, 1344-1348.

Stidham, S., 1985. Optimal control of admission to a queueing system. *IEEE Trans. Automatic Control*, AC 30, 705-713.

Taylor, J.W. 2004. “Volatility forecasting with smooth transition exponential smoothing”. *International Journal of Forecasting* 20, No. 2(Apr-Jun), 273– 286.

Taylor, J.W. 2006. “Invited Comments on “Exponential Smoothing: The state of the Art – Part II” by E.S. Gardner, Jr.” *International Journal of Forecasting* 22, No. 4(Oct-Dec), 671-672.

Theil, H. and Wage, H. 1964. Some observations on adaptive forecasting. *Management Science* 10, No. 2, 198-206.

van Ackere, A., 1995. Capacity management: pricing strategy, performance and the role of information. *Int. J. Production Economics* 40, No 1, 89–100

van Ackere, A. and Larsen, E.R. 2004. Self-organizing behavior in the presence of negative externalities: A conceptual model of commuter choice. *European Journal of Operational Research* 15, No. 2, 501-513

van Ackere, A.; Haxholdt, C.; and E.R. Larsen. 2006. Long and short term customer reaction: a two-stage queueing approach. *System Dynamics Review* 22, No. 4, 349-369.

Wolfram, S. 1994. *Cellular automata and complexity.* Westview Press, Champaign, IL.

Yechiali, U. 1969. On optimal balking rules and toll charges in the GI/M/1 queuing process. *Operations Research* 19, No. 2, 349-370.

Zohar, E. Mandelbaum, A. and N. Shimkin. 2002. Adaptive behavior of impatient customers in tele-queues: theory and empirical support. *Management Science* 48, No. 4, 566-583

AUTHOR BIOGRAPHIES

CARLOS A. DELGADO A. is a PhD student in Business Information Systems and teaching assistant at HEC Lausanne, the School of Business and Economics of the University of Lausanne, Switzerland. He received his Masters in Systems Engineering and a Bachelor of Industrial Engineering from the Faculty of Mines, National University of Colombia, where he also worked for three years as a Research Engineer and Teaching Assistant for the Bachelor course in systems simulation.

ANN VAN ACKERE is Professor of Decision Sciences at HEC Lausanne, the School of Business and Economics of the University of Lausanne, Switzerland since 1999. She obtained her PhD from the Stanford Graduate School of Business and joined the faculty of London Business School, UK, upon graduation.

KARTHIK SANKARANARAYANAN is a PhD student and research assistant at the Institute of Management, University of Lugano, Switzerland. He has a Masters in Embedded Systems Design and Management and a Bachelors degree in Electrical and Electronics Engineering. Previously he worked as a researcher at CEDT, Indian Institute of Science, Bangalore, India.

ERIK R. LARSEN is Professor at the Institute of Management and vice dean of the Faculty of Economics at the University of Lugano, Switzerland. Previously he held appointments at Cass Business School (London), London Business School, and Copenhagen Business School. During the period 1996-1998, he was an EU Marie Curie Fellow at the University of Bologna, Italy. He obtained his PhD from the Institute of Economics, Copenhagen Business School and his M.Sc. from the Technical University of Denmark.

Appendix D.

“Capacity adjustment in a service facility with reactive customers and delays: simulation and experimental analysis.”

Delgado, C. A., A. van Ackere, E. R. Larsen and S. Arango.

In *Proceedings of the 29th International Conference of the System Dynamics Society, Washington DC. July 2011*. Washington, DC: The System Dynamics Society, 2011.

CAPACITY ADJUSTMENT IN A SERVICE FACILITY WITH REACTIVE CUSTOMERS AND DELAYS: SIMULATION AND EXPERIMENTAL ANALYSIS.

Carlos Arturo Delgado Alvarez
HEC, School of Business and Economics, University of Lausanne
Dorigny, 1015-Lausanne, Switzerland
+41 21 692 34 67
E-mail: carlos.delgado@unil.ch

Ann van Ackere
HEC, School of Business and Economics, University of Lausanne
Dorigny, 1015-Lausanne, Switzerland
E-mail: ann.vanackere@unil.ch

Erik R. Larsen
Institute of Management University of Lugano
6904, Lugano, Switzerland
E-mail: erik.larsen@usi.ch

Santiago Arango
Complexity Center Ceiba, Faculty of Mines, Universidad Nacional de Colombia
Medellin, Colombia
E-mail: saarango@unal.edu.co

ABSTRACT

In this paper, we apply system dynamics to model a queuing system wherein the manager of a service facility adjusts capacity based on his perception of the queue size; while potential and current customers react to the managers' decisions. Current customers update their perception based on their own experience and decide whether to remain patronizing the facility, whereas potential customers estimate their expected waiting time through word of mouth and decide whether to join the facility or not. We simulate the model and analyze the evolution of the backlog of work and the available service capacity. Based on this analysis we propose two alternative decision rules to maximize the manager's cumulative profits. Then, we illustrate how we have developed an experiment to collect information about the way human subjects taking on the role of a manager in a lab environment face a situation in which they must adjust the capacity of a service facility.

KEYWORDS: Queuing system, capacity adjustment management, system dynamics, experimental economics, adaptive expectations

INTRODUCTION

Most typical research in queuing problems has been focused on the optimization of performance measures and the equilibrium analysis of a queuing system. Traditionally, analytical modeling and simulation have been the approaches used to deal with queuing problems. Most simulation models are stochastic and some more recent models are deterministic (van Ackere, Haxholdt, & Larsen, 2010).

The analytical approach describes mathematically the operating characteristics of the system in terms of the performance measures, usually in "steady state" (Albright & Winston, 2009). This method is useful for low-complexity problems whose analytical solution is not difficult to find. For complex problems, a simulation approach is preferable as it enables modeling the problem in a more realistic way, with fewer simplifying assumptions (Albright & Winston, 2009).

We consider those queuing systems in which customers decide whether or not to join a facility for service based on their perception of waiting time, while managers decide to adjust capacity based on their perception of the backlog of work (i.e. the number of customers waiting for service). The analysis of queuing problems could be aimed at either optimizing performance measures to improve the operating characteristics of a system or understanding how the manager and customers interact with the system to achieve their objectives. In the real world, queuing is a dynamic problem whose complexity, intensity and effects on the system change over time. Still, some problems may be modeled using the assumptions of classical queuing theory (Rapoport, Stein, Parco, & Seale, 2004). Considering the complexity of queuing problems, which is due to a set of interactive and dynamic decisions by the agents (i.e. customers and the manager) who take part in the system, we will focus on studying the behavioral aspects of queuing problems.

Haxholdt, Larsen, & van Ackere (2003) and van Ackere, Haxholdt, & Larsen, (2006); van Ackere et al., (2010) have applied deterministic simulation methodologies for studying behavioral aspects of a queuing system. Other authors have included cost allocation as a control for system congestion (queue size) (e.g. Dewan and Mendelson 1990). In this way, customers' decisions on whether or not to join the system are influenced by such costs. Likewise, those decisions can be based on steady-state (e.g. Dewan and Mendelson 1990). or be state-dependent (e.g. van Ackere 1995). The seminal papers on this subject are Naor (1969) and Yechiali (1971). Other authors have included dynamic feedback processes to build perceptions of the behavior of the queue (van Ackere et al., 2006) and/or of demand (van Ackere et al. 2010), which influence the decisions of customers and managers. A more detailed discussion of the state of the art on behavioral aspects in queuing theory can be found in (van Ackere et al., 2010).

We propose two methodological approaches to achieve our goals. Firstly, we use system dynamics to learn about the macro-dynamics of customers and the manager interacting in a service facility. Specifically we analyze how the available service capacity and the queue evolve and how the delay structure affects the manager’s decision. We also want to assess how the manager adjusts capacity based on the evolution of the backlog of work (i.e. the number of customers waiting for service). Haxholdt et al. (2003) and van Ackere et al. (2006 and 2010) applied system dynamics to tackled similar problems. System dynamics is useful for problems, which do not require much detail. That is, those which can be modeled at a high level of abstraction. This kind of problems is usually situated at the macro or strategic level (e.g. marketplace & competition, population dynamics and ecosystem) (Borshchev & Filippov, 2004)

Next we apply experimental economics (Smith, 1982) to capture information about how subjects playing the role of a manager in a lab environment, decide when and by how much to adjust the capacity of a service facility. We use the system dynamics based simulation model as a computational platform to perform the experiment. For more details about how system dynamics models have been used to carry out laboratory experiments, see (Arango, Castaneda, & Olaya, 2011). Experimental economics is a methodology that based on collecting data from human subjects to study their behavior in a controlled economic environment (Friedman & Sunder, 1994).

This paper is organized as follows: Firstly, we discuss the dynamic hypothesis of the problem proposed initially by van Ackere et al. (2010) and explain why we modify the model. Then, we analyze the model behavior of the base case. In the following section, we introduce two alternative strategies to manage the capacity adjustment of the service facility. We determine the optimal parameters for these strategies and analyze the resulting system behavior. We also perform a sensitivity analysis to the parameter values. Finally, we present the experimental laboratory and discuss the collected results.

A SERVICE FACILITY MANAGEMENT MODEL

In this section, we analyze the dynamic hypothesis of the queuing model proposed by van Ackere et al. (2010). This model captures the relationship between customers and manager (referred to as the service provider) as agents who interact in a service system. The causal loop diagram of Figure 1 portrays the feedback structure of these two actors in the system. The model consists of two sectors: the customers' behavior is to the left and that of the manager to the right. Both sectors are connected by the queue, whose evolution determines the dynamics of these actors in the system. Customers decide whether to use the facility based on their estimate of waiting time, while the manager decides to adjust the service capacity based on the queue length. Examples of this kind of system include a garage where customers take their car for maintenance, and workers or students who daily patronize a restaurant to have lunch. In both examples, customers are free to use or not the facility for service and the manager is motivated to encourage customers to use his facility by adjusting its service capacity.

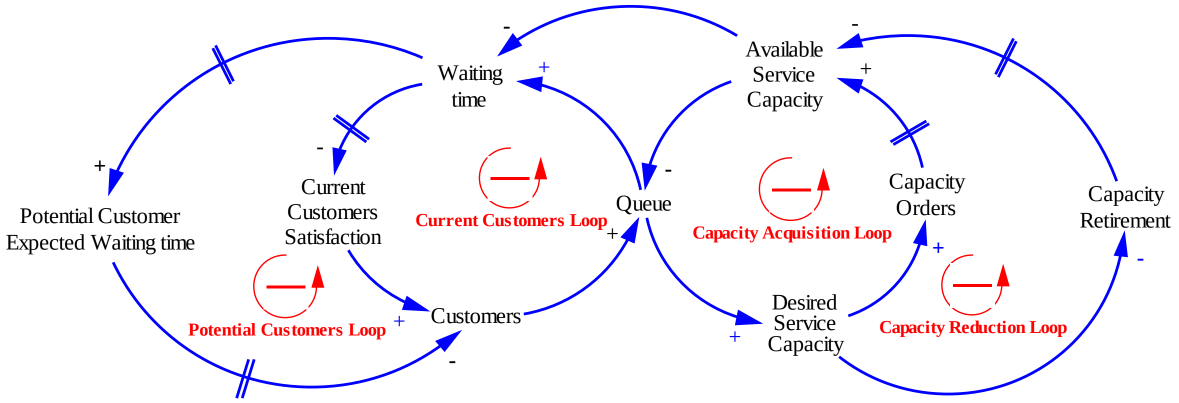


Figure 1. Feedback loop structure for a customers-facility queuing system

Two groups of customers are assumed: current and potential customers. The former make up the customer base of the facility; they periodically patronize it as long as they are satisfied. They consider being satisfied when their expected waiting time is less than the market reference, which they find acceptable. The second group represents those customers who the manager envisages as potentially attractive to the business. They can be either former customers, who left due to dissatisfaction, or new customers who require the service and look for a facility. They decide whether or not to join the facility depending on their expected waiting time, which they also compare to the market reference.

Customers form their perception of waiting time ($\dot{W}_t$) each period using adaptive expectations (Nerlove, 1958), as shown in Equation 1:

$$\dot{W}_t = \varphi * W_{t-1} + (1-\varphi) * \dot{W}_{t-1} \qquad (1)$$

where φ is called the coefficient of expectations (Nerlove, 1958) and $1/\varphi$ may be considered as the time taken by customers to adapt their expectations. Current customers adjust their expectation based on their last experience (W_t), while potential customers rely on word of mouth. The decision of joining a facility for service based on its reputation often requires more time than when we base this decision on our own experience. Thus, we assume that the time required by potential customers to adapt their expectations is longer than or equal to that of the current customers.

While the current customers' perception determines their loyalty to the facility, the potential customers' perception defines if they will join the customer base. The lower the waiting time perceived by current customers, the more loyal they are, whereas the higher the perceived waiting time, the more customers will leave the customer base. Regarding potential customers, the lower their expected waiting time, the more will become new customers for the facility. The rates at which new customers join the customer base and current customers leave it are modeled using nonlinear functions of the satisfaction level. van Ackere et al. (2010) discuss some alternatives to model these functions.

To summarize the customers' dynamics: longer queues bring about higher waiting times for current customers and increased perceptions of waiting time for potential customers, implying that the level of satisfaction with the facility's service of both customer groups decreases. Consequently, over time this reduction in customers' satisfaction leads current customers to leave the facility and discourages new customers from joining it in the future. Thus, the number of customers waiting for service will decrease until the waiting time tends to acceptable levels compared to the market reference and the customers' perception stabilizes. These dynamics are described by the two balancing loops to the left in Figure 1.

As far as the service provider (the right side of Figure 1) is concerned, van Ackere et al. (2010) model the type of service systems where the capacity adjustment involves an implementation time. For instance, hiring new employees requires new training, laying off staff may imply a notice period, acquiring new IT systems takes time, among others. However, the authors represent this time in the model using an information delay (Sterman, 2000); after the manager estimates the required capacity, any needed adjustment is implemented gradually. This is a simplified view of the delay structure. In a system dynamics context, this kind of delays is better modeled through material delays, which capture the real physical flow of the capacity (Sterman, 2000). Once the adjustment decision has been made, its implementation process does not materialize immediately. We deviate from van Ackere et al. (2010) by incorporating this material delay structure in the model, as the stock and flow diagram of Figure 2 illustrates. In this way, we can model how the manager accounts for his previous decisions, which have not yet taken effect, to make his next decision.

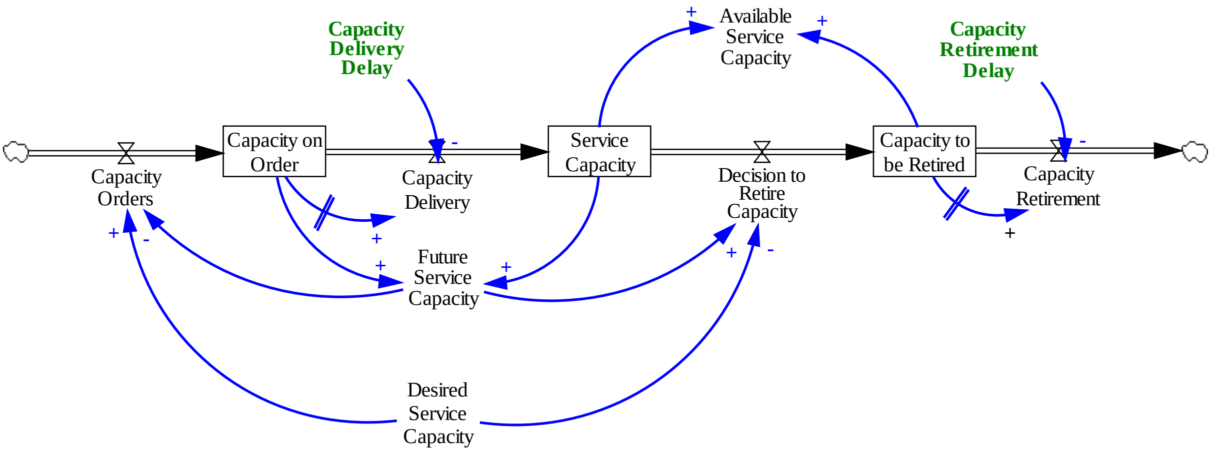


Figure 2. System dynamics representation for the capacity adjustment management of a service facility.

The capacity adjustment process is depicted in Figure 2 by capacity orders and the decision to retire capacity, which determine the available service capacity. Starting from the left, the manager decides how fast and how much to adjust capacity based on his desired service capacity and the future capacity. The latter is explained below and depends on his previous decisions. He estimates the desired service capacity based on his perception of the average queue length and a market reference for the waiting time (τ_{MR}). Like the customers, the manager forms this perception by applying adaptive expectations. He updates his expected average queue length based on the most recent observation of the queue (Q_{t-1}). This expected average queue length (EQ_t) is given by:

$$EQ_t = \beta * Q_{t-1} + (1-\beta) * EQ_{t-1} \qquad (2)$$

where β is the coefficient of expectations for the manager and $1/\beta$ may be interpreted as the time required by the manager to adapt his perception. Then, the desired service capacity of the manager is determined as follows:

$$DC_t = \frac{EQ_t}{\tau_{MR}} \qquad (3)$$

The longer the queue the greater the desired service capacity and the larger the capacity orders (c.f Figure 1). After the manager decides how much capacity to add (c.f. capacity orders in Figure 2), these orders accumulate as capacity on order (*CO*) until they are available for delivery (c.f. capacity delivery delay in Figure 2). Some examples of this kind of delayed process in capacity acquisition include construction of new buildings, purchase of new equipment and hiring staff. Once the capacity order is fulfilled, the service capacity (*SC*) will be increased by the capacity delivery. The greater the service capacity, the higher the service rate and thus fewer customers waiting. In this way, a third balancing loop (c.f. capacity acquisition loop in Figure 1) results from the dynamics between the manager and customers.

The decision of adjusting capacity may also imply removing capacity. When this occurs, the capacity, which the manager decides to withdraw, will be designated as capacity to be retired (CbR). This capacity remains available to the customer during the capacity retirement delay (e.g. end a lease on a building, notice period for staff, etc). Hence, the currently available service capacity at the facility at time *t* is given by,

$$\mathrm{ASC}_t = \mathrm{SC}_t + \mathrm{CbR}_t \tag{4}$$

After the delay involved in the capacity retirement, the available service capacity will decrease due to this retirement, as shown in Figure 1, and the number of customers in the queue will thus increase. This effect yields the fourth balancing loop in the system. This loop describes the behavior caused by the decisions of capacity reduction.

Finally, the capacity that will be available once all the manager's decisions have been implemented, i.e. the future capacity, is given by,

$$\mathrm{FSC}_t = \mathrm{CO}_t + \mathrm{SC}_t \tag{5}$$

Then, Equations (4) implies that FSC_t equals

$$\mathrm{FSC}_t = \mathrm{ASC}_t + \mathrm{CO}_t - \mathrm{CbR}_t \tag{6}$$

To summarize the manager's dynamics: longer queues increase his desired service capacity. The higher this desired service capacity, the more capacity the manager orders or the less he removes. Over time, the capacity orders will increase the available service capacity, while the capacity retirement will decrease it. Consequently, the higher (the lower) the available service capacity the lower (the higher) the number of customers queuing. Like the customers' dynamics, the two balancing loops, which describe the manager's behavior, may lead to stabilizing his perception over time. Thus, we are interested in studying how the manager analyzes the customers' behavior in order to adjust capacity and how the multiple delays involved in the system affect his decisions.

MODEL BEHAVIOR

Before trying out some alternative policies or strategies to model the manager's decisions and discussing descriptively some experimental results, we analyze the typical behavior of the system occurring when one of the equilibrium conditions is modified. The model is initially set under the equilibrium conditions, which are described in Table 1. Then we illustrate the impact on the system behavior of increasing the size of the initial customer base from 175 to 200. The other initial values remain as shown in Table 1. We simulate the model for 100 time units using a simulation step of 0.0625 time units.

State Variables	**Equilibrium Value**	**Unit**
Customer base	175	People
Queue	50	People
Average queue	50	People
Capacity on order	0	People / Time
Service capacity	25	People / Time
Capacity to be retired	0	People / Time
Perceived waiting time of current customer	2	Time unit
Perceived waiting time of potential customers	2	Time unit
Exogenous Variables	**Value**	**Unit**
Visit per time unit	0.15	1 / Time unit
Market reference waiting time (τ_{MR})	2	Time unit
Delays	**Value**	**Unit**
Time to perceive queue length (1 / β)	4	Time unit
Capacity delivery delay	4	Time unit
Capacity retirement delay	2	Time unit
Perception time of current customers (1 / φ_c)	2	Time unit
Perception time of potential customers (1 / φ_p)	4	Time unit

Table 1. Initial conditions of equilibrium

Figure 3 illustrates the evolution of the available service capacity and the number of customers waiting for service. We can observe that the manager adjusts the service capacity by imitating the evolution of the queue (i.e. the backlog of work). In this sense, he is trying to keep the average waiting time close to the market reference and while keeping the utilization rate close to 1, as shown in Figure 4. The lags involved in the manager and customer dynamics in addition to the manager's reaction result in the oscillating phenomenon and a certain decreasing tendency, as shown in Figure 3. Next, we go into more detail of the causes of this pattern.

An increase in the customer base will raise the arrival rate. Considering that the service capacity remains constant due to the lags involved in the capacity adjustment process and the formation of perceptions by the manager, more customers will wait for service. As the queue increases, the manager adjusts gradually his desired service capacity. According to Figure 1, the higher the desired service capacity, the larger the capacity orders. However, the capacity is delivered after 4 periods. The average waiting time therefore increases initially as plotted in Figure 4, affecting the perception of current customers and the expected waiting time of potential customers. When the perception of waiting time exceeds the market reference (2 time units), the customer base starts to decrease because more current customers are dissatisfied and fewer potential customers wish to join the facility. Hence, when the manager's decisions to add capacity start to materialize, the backlog of work (i.e. the queue) is falling. Consequently, the available service capacity reaches its peak at about the time the queue is reaching its nadir. Moreover, the manager reacts again to this behavior of the customers, but on this occasion by reducing his available service capacity to avoid having idle capacity. Neither manager nor customers consider the delays inherent in the reaction of each other. Hence, the backlog soars because of the manager's decision. Thus, despite the manager trying to adjust the service capacity by imitating the evolution of the queue, the multiple delays in the system bring about a fluctuating pattern as illustrated in figure 3.

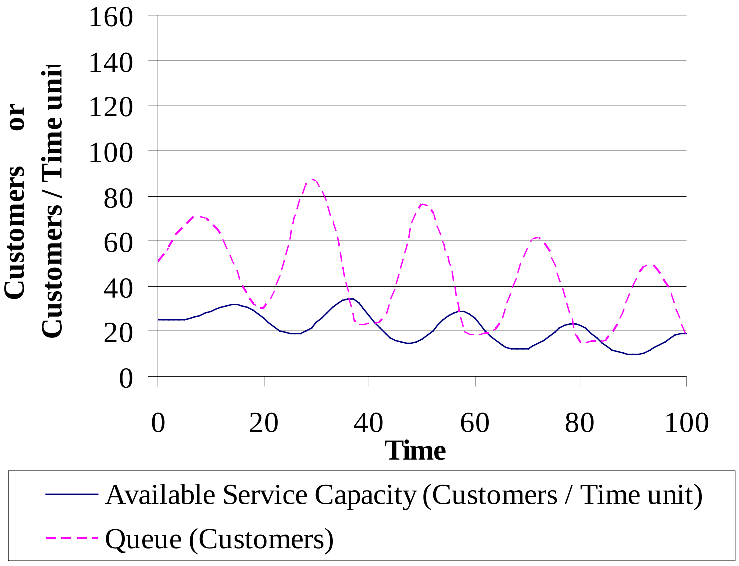
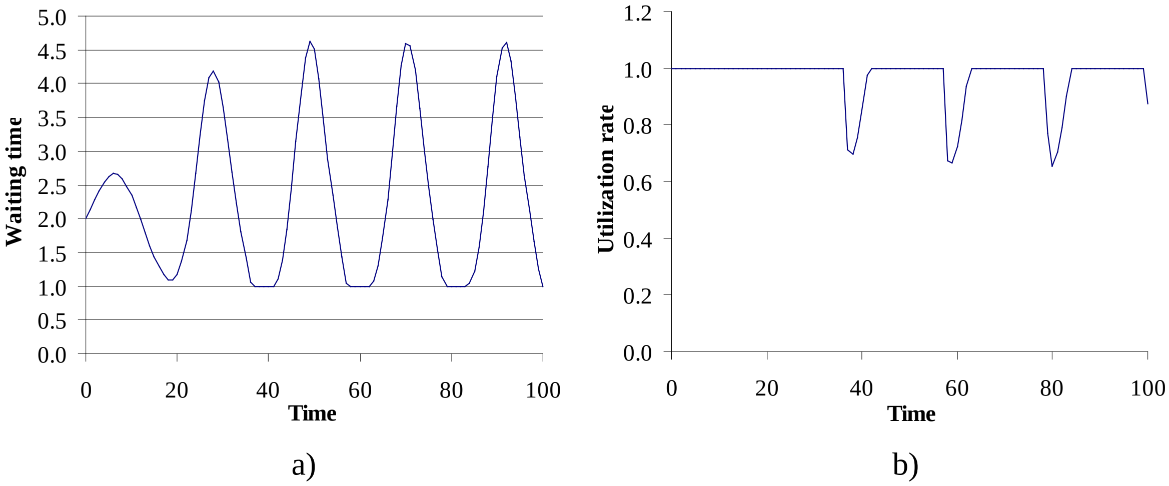


Figure 3. Illustrative behavior of the available service capacity and queue length



a) b)

Figure 4. Illustrative behavior of (a) the average waiting time and (b) the utilization rate.

We have explained the model and illustrated a typical case where the manager reacts to customers' dynamics. In the next section, we propose other alternative decision rules to enable the manager to adjust capacity more effectively. These rules are based on the manager's perception of the backlog of work. Two alternative ways to form this perception based on the evolution of the queue are introduced. The decision rules consider both the required capacity adjustment and the speed at which this adjustment is carried out.

ALTERNATIVE DECISION RULES

The aim of the manager is to maintain sufficient available service capacity (ASC_t) in his facility in order to satisfy the customers. He thus decides whether to adjust the service capacity and at what time to do so. We propose a heuristic to determine the required capacity adjustment (RCA_t) by incorporating the speed at which the manager decides to adjust it. Let α be the service provider's speed to adjust capacity, i.e. how fast he decides to either add or reduce capacity. We defined above that the capacity adjustment decisions depend on the future service capacity (FSC_t), and the desired capacity (DC_t). Thus, including α in this definition, we may state RCA_t as follows:

$$RCA_t = \alpha * (DC_t - FSC_t), \qquad (7)$$

where α must be nonnegative and less than 1. This adjustment involves either an increase in capacity (when DC_t - FSC_t >0), a decrease in capacity (when DC_t - FSC_t < 0), or leaving capacity unchanged (when DC_t - FSC_t = 0). Taking into account that the capacity delivery delay may be different from the capacity retirement delay (c.f. Figure 2), we assume that the speed to either add or remove capacity can also be different. In this sense, the parameter α is determined as follows:

$$\alpha = \begin{cases} \alpha_1 & \text{if } DC_t - FSC_t < 0 \\ \alpha_2 & \text{if } DC_t - FSC_t >= 0 \end{cases} \qquad (8)$$

where DC_t and FSC_t are as defined in Equation 3 and 6. Consider now that the manager does not necessarily keep in mind all his previous decisions, some of which are still in the process of execution. Thus, the future service capacity (FSC_t), which the manager perceives, would be modeled as:

$$\mathrm{FSC_t = ASC_t} + \gamma * \mathrm{(CO_t - CbR_t)} \qquad (9)$$

where γ represents the proportion of the capacity adjustment that has not yet been implemented, which the manager takes into account. Replacing α, DC_t and FSC_t using Equations 8, 3 and 9, respectively, in Equation 7, the decision of how much to adjust capacity each period is determined by

$$RCA_t = \alpha * \left(\frac{\beta * Q_{t-1} + (1-\beta) EQ_{t-1}}{\tau_{MR}} - ASC_t - \gamma * (CO_t - CbR_t) \right) \qquad (10)$$

s.t.

$$\alpha = \begin{cases} \alpha_1 & \text{if} \quad \dfrac{\beta * Q_{t-1} + (1-\beta)EQ_{t-1}}{\tau_{MR}} - ASC_t - \gamma * (CO_t - CbR_t) < 0 \\ \alpha_2 & \text{if} \quad \dfrac{\beta * Q_{t-1} + (1-\beta)EQ_{t-1}}{\tau_{MR}} - ASC_t - \gamma * (CO_t - CbR_t) >= 0 \end{cases} \tag{11}$$

We propose a second manner to estimate *DC*$_t$. Instead of using adaptive expectations, the manager may simply consider the most recent backlog, i.e. customers waiting for service (Q_t), to estimate demand. That is, he looks at his current order book to decide how much capacity is required. Such an attitude is meaningful in situations where capacity can be adjusted fairly cheaply and quickly, e.g. by using temporary staff. In this case Equations 10 and 11 become:

$$RCA_t = \alpha * \left(\frac{Q_t}{\tau_{MR}} - ASC_t - \gamma * (CO_t - CbR_t) \right) \tag{12}$$

s.t.

$$\alpha = \begin{cases} \alpha_1 & \text{if} \quad \dfrac{Q_t}{\tau_{MR}} - ASC_t - \gamma * (CO_t - CbR_t) < 0 \\ \alpha_2 & \text{if} \quad \dfrac{Q_t}{\tau_{MR}} - ASC_t - \gamma * (CO_t - CbR_t) >= 0 \end{cases} \tag{13}$$

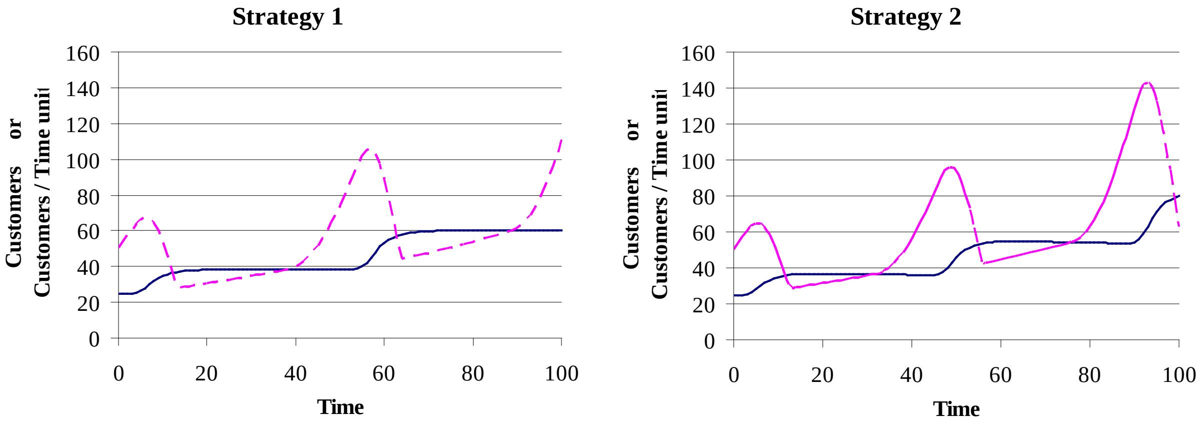
Optimal Strategies

Our objective is to find optimal values for the parameters $\alpha_1, \alpha_2, \beta$ and γ, which determine the above two strategies, to maximize the manager's cumulative profits over 100 time units. In order to calculate this profit we introduce a fixed cost and revenue resulting from providing the service. The equations 10 to 13 are nonlinear and thus complicated to optimize analytically. Thus, we apply simulation optimization (Keloharju & Wolstenholme, 1989; Moxnes, 2005) in order to find the optimal parameter values.

We use the optimizer toolkit of Vensim where the cumulative profits are set as the payoff function. The optimal parameter values we obtain are given in Table 2. According to this table, the second strategy, i.e. when the manager forms his perception based on the most recent value of the backlog, reaches the best payoff (2'151 compared to 2'059 for strategy 1). This occurs because when using strategy 2 the manager makes decisions a bit more aggressively than when using strategy 1, as shown figure 5. Hence, the manager reaches higher profits when he relies on the most recent information about the customers' behavior, i.e. Q_t. The optimal value of β (i.e. the coefficient of expectations), which equals 1 (see table 2), for strategy 1 strengthens the above remark. A coefficient of expectation equal to 1 means that the manager updates his expectation by using only the most recent information regarding the backlog. That is, the manager does not account for the past. In that case, Q_{t-1} is the latest information about the backlog the manager has to update his perception, EQ_t, at time unit *t*.

Strategy	Alpha 1	Alpha 2	Beta	Gamma	Maximum Payoff Value
Adaptive expectations	1.00	0.00	1.00	0.40	1'950
Most recent value of the backlog	1.00	0.00	N.A	0.37	2'071

Table 2. Optimal values of the parameters which define each strategy



(a)

(b)

Figure 5. Evolution of the queue (i.e. backlog) and the available service capacity for the two capacity adjustment strategies with the optimal parameter values.

Figure 5 shows the behavior of the two parts of the system (customers and the manager) for both strategies. Their optimal behaviors are similar. Like in the base case, when the manager applies either of these two strategies, the backlog grows at the beginning of the simulation and the manager reacts by increasing capacity. However, as he bases his decisions on the most recent information about the backlog, he notices quickly that the backlog goes down. Thus, his decision to increase capacity becomes less aggressive resulting in the utilization rate gradually increasing back to 1 (see Figure 6). Consequently, the manager's decisions encourage current customers to remain loyal which in turn encourages the manager to keep the available service capacity constant. The manager's behavior brings about current customers being satisfied and thus inducing potential customers to patronize the facility through word of mouth. New customers joining the customer base imply that the arrival rate steeply increases. The manager responds by slowly increasing the available service capacity, which quickly reduces the queue. From this point onwards, an oscillating phenomenon starts to emerge. This oscillating pattern differs from that of the base case in that it grows exponentially over time.

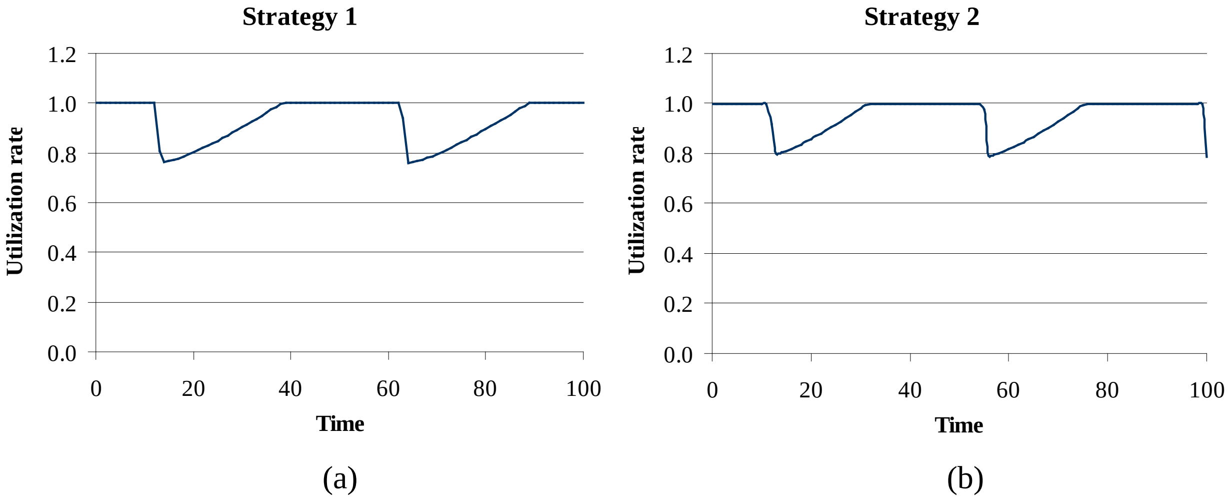


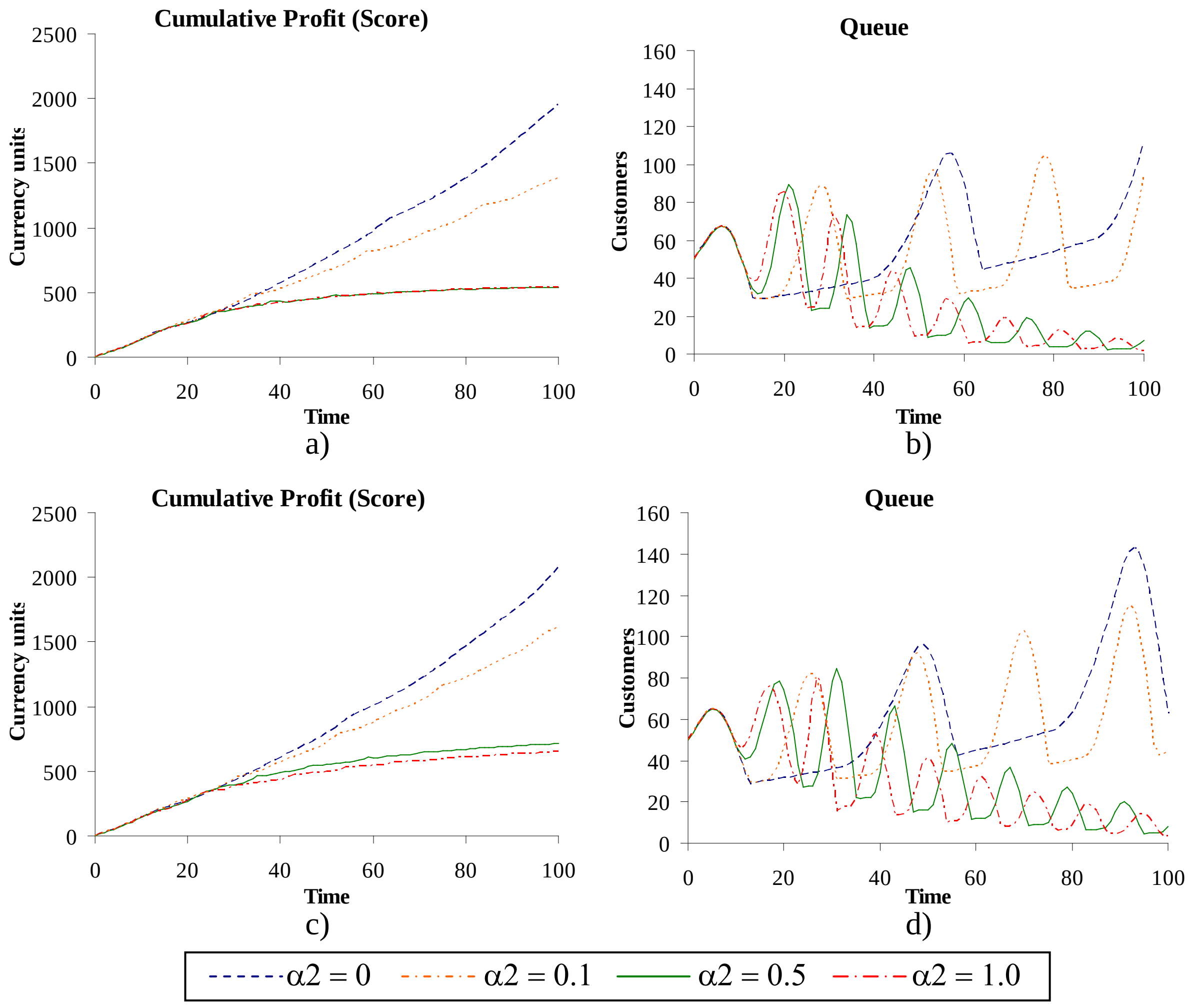
Figure 6. Evolution of the utilization rate for the two capacity adjustment strategies set up with the optimal parameter values.

Sensitivity analysis

We perform a sensitivity analysis to understand the impact of the different parameters, which define the alternative strategies, on the model behavior. In particular, we analyze the effect of a change in the values of these parameters on the manager's cumulative profits and the evolution of the queue.

First we illustrate the case in which we change α_2 (i.e. the speed at which the manager removes capacity). We select this parameter because it has the strongest impact. Figure 7 illustrates how changing the value of α_2 in both strategies affects the evolution of the queue and the manager's cumulative profits. We can observe that changes in these two variables emerge after about 27 time units, particularly, when α_2 is large (e.g. 0.5 or 1.0), i.e. when the manager quickly removes capacity. For instance, using both strategies with α_2 equal to 1.0 the cumulative profits decrease about 70% compared to the optimal value, while the backlog decreases by about 98% for strategy 1 and 94% for strategy 2. Likewise, the higher the parameter, the more the backlog oscillates.

Changes in the other parameters have small impacts on the evolution of the cumulative profits and the queue. As far as α_1 (i.e. the speed at which the manager add capacity) is concerned, for very small values (e.g. 0.0 and 0.1) the manager's cumulative profits and the queue are slightly reduced using both strategies. Regarding the speed at which the manager updates his perception in Strategy 1, i.e. β, varying this parameter results in similar effects as changing α_1. Finally, by trying different values of γ we found that they do not have any significant impact.



a)

b)

c)

d)

Figure 7. The cumulative profits and queue length when strategies 1 (Figs a and b) and 2 (Figs c and d) are simulated for selected values of α_2, keeping values of α_1, β, and γ constant as shown in Table 2.

A SERVICE FACILITY MANAGEMENT EXPERIMENT

We use the model described above as a computational platform to implement a laboratory experiment (c.f. Smith, 1982). The objective behind this experiment is to collect experimental information to assess how human subjects taking on the role of a manager face a situation in which they must adjust the capacity of a service facility. We also want to analyze how they use the available information to make capacity adjustment decisions. The subjects have information about the behavior of both the facility and the customers. Regarding the facility, they know the past and current available service capacity and utilization rate. As for customers, subjects know the past and current backlog (i.e. the number of customers waiting for service).

Experimental Protocol

We design this experiment based on the protocol for experimental economics (e.g. Smith, 1982; Friedman and Sunder 1994). We recruited undergraduate and master students in Finance, Management and Economics from the University of Lausanne. They were invited to participate in an experiment designed to study decision making in a service industry, through which they could earn up to 80 Swiss Francs. We received about 400 replies and selected 187 subjects following the principle of "first come, first served" in order to perform six experimental treatments. Each treatment had at least 30 participants. Subjects were allocated across eleven experimental sessions; each involved around 16 subjects and lasted, on average 90 minutes. Two facilitators supervised each session. The task of the subjects was to use a computer based interface, which portrayed the service capacity adjustment problem of a garage, to decide each period how much capacity to add or remove. They had to perform this task for 100 experimental periods.

This experiment was conducted in the informatics laboratories of the School of Business and Economics. Upon arrival at the laboratory, the subjects were allocated to a PC and separated from their neighbor by another PC. Communication between the subjects was forbidden. Once they were seated, we gave them written instructions and a consent form, which they had to sign before starting the experiment. Then, a short introduction to the experiment was presented to them. The instructions were quite simple and provided subjects with a short explanation of the system that they had to manage in the experiment and all the information, which they had available to carry out their task. We present the instructions and the interface used to run the experiment in the appendix of this paper.

We gave the subjects the payoff scale through which they earned their reward depending on their performance in the experiment. Performance was measured based on the cumulative profits that subjects had at the end of the experiment, i.e. at the period 100 or when the available service capacity reached 0, If that happened before than the period 100.

Experimental Treatments

In addition to the base case, we have designed other five experimental treatments to understand how the manager adjusts the capacity of an industry service. These five treatments are divided in two groups to study the effect of different factors. The first group is composed of four treatments and its objective is to analyze how the delay structure, inherent to the system, affects how the manager decides to adjust capacity. This delay structure includes the delays the manager knows (i.e. the implicit lags in capacity adjustment), and those which are unknown to him (i.e. the time required by potential and current customers to update their perceptions). The last group has a single treatment, which includes a cost to add or remove capacity. Table 3 summarizes the conditions of each treatment.

Treatment	Current customers Delay	Potential customers Delay	Time to increase capacity	Time to decrease capacity	Cost per unit change in capacity
Base Case	4	2	4	2	-
A	10	2	4	2	-
B	6	4	4	2	-
C	4	2	8	4	-
D	4	2	2	1	-
E	4	2	4	2	1

Table 3. Treatment conditions.

EXPERIMENTAL RESULTS

All subjects overreact to the initial increase of the backlog. This sudden rise is independent of subjects' decisions since it depends on the initial conditions. Thus, we can interpret this first reaction of the subjects as a learning process in which they are trying to adapt to the system behavior. In other words, we can call this initial period a transition period. Recall that we observed a similar pattern of the backlog in the simulation results.

From this point onwards, we identify three groups of subjects, whose decisions result in similar behavioral patterns. Figure 8 illustrates the evolution of the backlog and the available service capacity of two typical subjects of each group. The first group is composed of those subjects who overreact strongly to the initial overshoot of the backlog and then they make many small decisions to gradually adjust capacity over time (e.g. Subjects 5 and 11). Most of these decisions concern capacity addition. Consequently, the garage's available service capacity for this kind of managers presents an exponential increase over time. After the initial transition, the available service capacity and the queue behave in the same way. Thus, we can consider that these subjects quickly learn to manage the system to achieve sustainable growth. The subjects in this group achieved the higher scores of the experiment.

The second group (e.g. Subjects 12 and 18) represents those subjects who, after their slight overreaction to the initial backlog, make fewer but more aggressive capacity adjustment decisions than the subjects of the first group. Moreover, they continue to overreact to the evolution of the backlog over time. This behavior results in an oscillating pattern for both the backlog and the available service capacity: they increase exponentially, but more slowly than for the first group. These two groups, despite achieving quite different behavioral patterns compared to the two optimal strategies discussed before, attain similar total profits.

The last group includes subjects who, even after the transition period, continue to overreact significantly to the evolution of the backlog (e.g. Subjects 3 and 30). Although in some cases the backlog evolves as when simulating the optimal strategies (see Figure 5), the subjects did not capture the customers' behavior. We can consider that these subjects were unable to handle the delay structure inherent to the system. They performed poorly, achieving the lower payoffs, and occasionally finding themselves with zero service capacity before the end of the experiment.

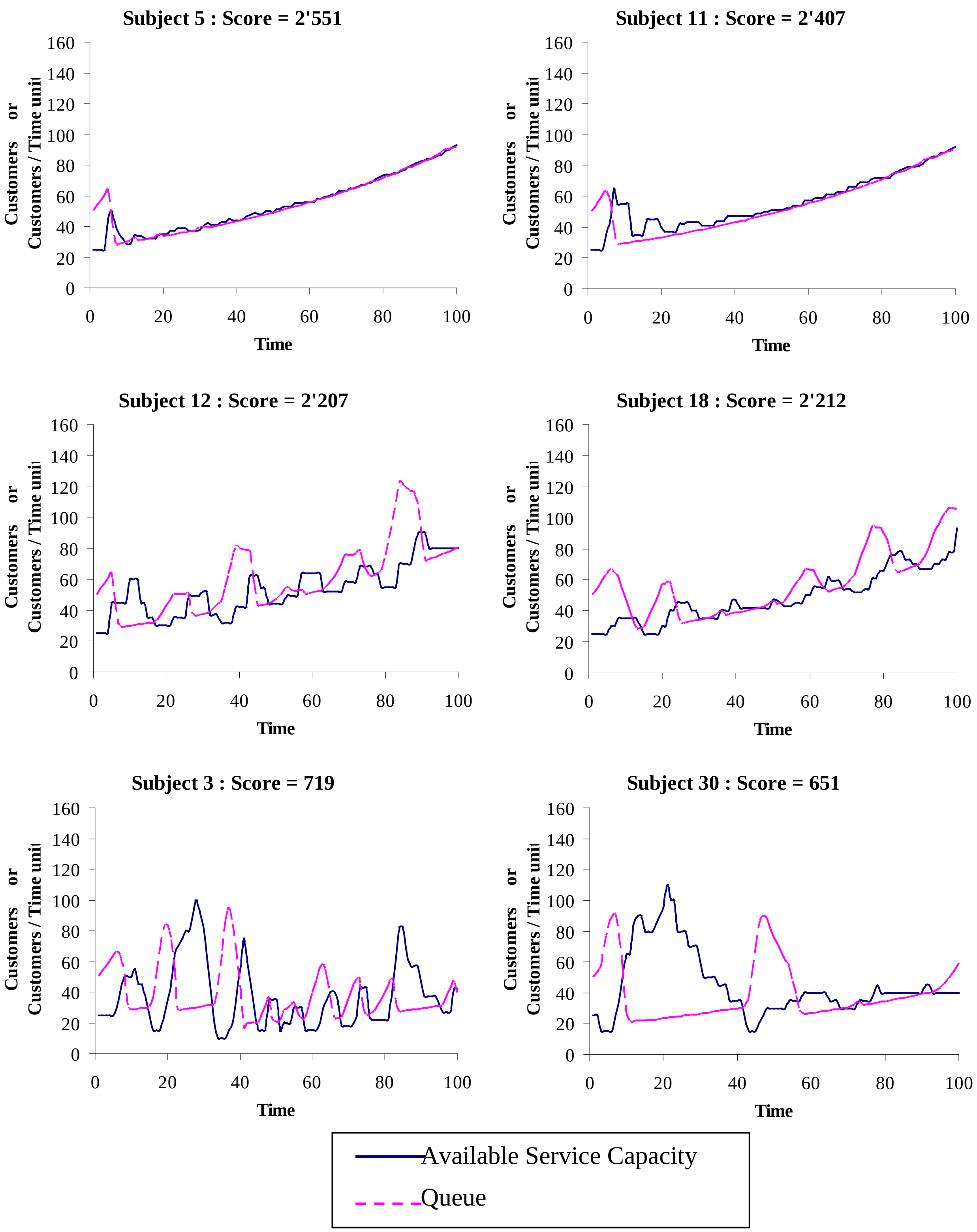


Figure 8. Experimental results for six typical subjects.

Treatment Results

The outcomes of the treatments were compared using the *Wilcoxon Rank-Sum test or Mann-Whitney U test*. Table 4 shows the corresponding p-values. Using a 0.05 significance level, these p-values enable us to interpret that the cumulative profits achieved in treatments C (i.e., slow adjustment) and D (i.e., fast adjustment) are, on average, significantly different compared to the cumulative profits achieved in the other treatments. By looking at the box plots in Figure 9 we can get an idea of such a difference as the mean cumulative profits of treatments C and D are either above or below the mean cumulative profits of the other treatments, supporting the remark inferred from the Wilcoxon Rank-Sum tests. We can also observe that the variability in treatment D is less compared to that of the other treatments. In addition, the distributions of treatments A, C, D and E are reasonably more symmetric than those of treatment B and the base case.

Col Mean - Row Mean P-Values	Basecase	Treatment A	Treatment B	Treatment C	Treatment D
Treatment A	0.2805				
Treatment B	0.9035	0.1772			
Treatment C	0.0008	0.0029	0.0000		
Treatment D	0.0002	0.0000	0.0003	0.0000	
Treatment E	0.2871	0.7562	0.2310	0.0003	0.0000

Table 4. P-values of the Wilcoxon Rank-Sum test for the cumulative profits

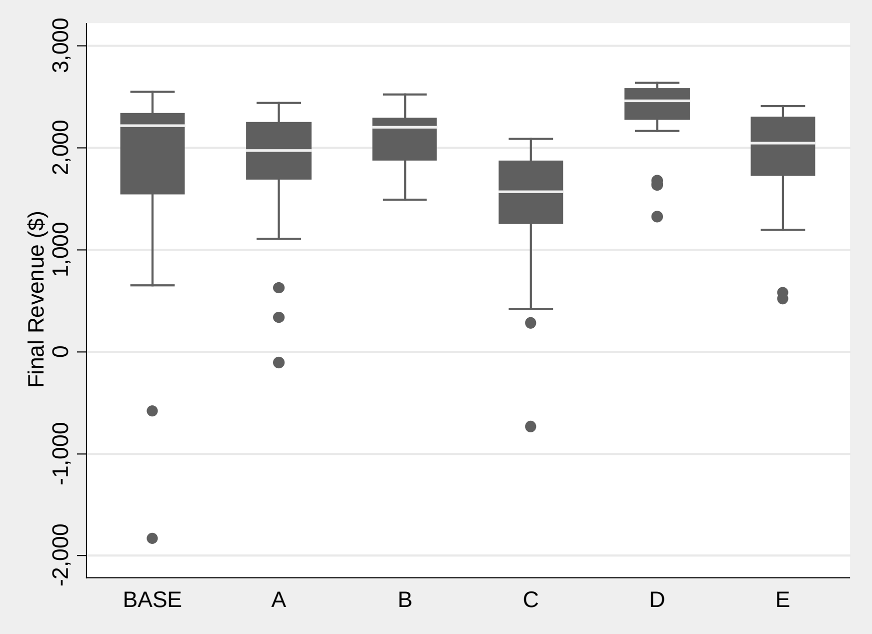


Figure 9. Box plots for the cumulative profits by treatment

CONCLUSIONS AND FURTHER WORK

In this paper, we have applied a system dynamics model to study how the manager of a service facility adjusts capacity based on his perception of the queue length, whereas potential and current customers react to the managers' decisions. While current customers update their perception based on their own experience and decide whether to stay in the customer base, potential customers update their perception through word of mouth and decide whether to join the customer base. We have simulated the model and analyzed the evolution of the backlog of work and the available service capacity. Based on this analysis we have proposed two alternative decision rules to maximize the manager's cumulative profits. Then, we have illustrated how we developed an experiment to collect information about how human subjects taking on the role of a manager in a lab environment face a situation in which they must adjust the capacity of a service facility.

Simulating this queuing model showed that when the manager tries to adjust the service capacity by imitating the evolution of the queue (i.e. the backlog of work), the multiple delays in the system bring about an oscillatory phenomenon. Optimizing the parameters, which set the alternative strategies, we found that the manager reaches higher profits when he relies on the most recent information about the customers' behavior, i.e. the most recent backlog. The sensitivity analysis enables to conclude that changes in the speed at which the manager removes capacity have a strong impact on the evolution of the available service capacity and the backlog. Varying the other parameters results in small impacts on the evolution of these two variables.

As far as the experiment is concerned, we identify three groups of subjects, whose decisions bring about similar behavioral patterns. The first group included the subjects who overreact strongly to the initial sudden increase of the backlog and make many small decisions to gradually adjust capacity over time. The second group represented the subjects who, after overreact to the initial backlog slightly, they make fewer but more aggressive capacity adjustment decisions than the subject of the first group. The last group included subjects who even, after the transition period, overreact significantly to the backlog. The two first groups, despite quite different behavioral patterns compared to the two optimal strategies discussed, achieved similar total profits.

The next step will be estimate a decision rule which adjusts to collecting data from Subjects. Extensions include incorporating prices to manager' decisions, i.e. a unit cost for each unit of capacity which the manager decides to add or remove. An interesting approach would be to conduct another experiment wherein another group of human subjects will assume the role of customers.

REFERENCES

Albright, S. C., & Winston, W. L. (2009). *Management Science Modeling* (Revised 3r., p. 992). South Western Cengage Learning.

Arango, S., Castaneda, J., & Olaya, Y. (2011). Laboratory Experiments and System Dynamics. Medellin.

Borshchev, A., & Filippov, A. (2004). From system dynamics and discrete event to practical agent based modeling: reasons, techniques, tools. *The 22nd International Conference of the System Dynamics Society*. Oxford.

Dewan, S., & Mendelson, H. (1990). User Delay Costs and Internal Pricing for a Service Facility. *Management Science*, *36*(12), 1502-1517.

Friedman, D., & Sunder, S. (1994). *Experimental methods: a primer for economists* (1st ed.). Cambridge: Cambridge University Press.

Haxholdt, C., Larsen, E. R., & van Ackere, A. (2003). Mode Locking and Chaos in a Deterministic Queueing Model with Feedback. *Management Science*, *49*(6), 816-830.

Keloharju, R., & Wolstenholme, E. F. (1989). A Case Study in System Dynamics Optimization. *Journal of the Operational Research Society*, *40*(3), 221-230. doi:10.1057/palgrave.jors.0400303

Moxnes, E. (2005). Policy sensitivity analysis: simple versus complex fishery models. *System Dynamics Review*, *21*(2), 123-145. doi:10.1002/sdr.311

Naor, P. (1969). The Regulation of Queue Size by Levying Tolls. *Econometrica*, *37*(1), 15-24.

Nerlove, M. (1958). Expectations and Cobweb Phenomena. *The Quarterly Journal of Economics*, *72*(2), 227-240.

Rapoport, A., Stein, W. E., Parco, J. E., & Seale, D. A. (2004). Equilibrium play in single-server queues with endogenously determined arrival times. *Journal of Economic Behavior & Organization*, *55*(1), 67-91. doi:10.1016/j.jebo.2003.07.003

Smith, V. L. (1982). Microeconomic Systems as an Experimental Science. *The American Economic Review*, *72*(5), 923-955.

Sterman, J. D. (2000). *Business Dynamics: Systems thinking and modeling for a complex world* (p. 982). Chicago, IL: Irwin-McGraw Hill.

van Ackere, A. (1995). Capacity management: Pricing strategy, performance and the role of information. *International Journal of Production Economics*, *40*(1), 89-100.

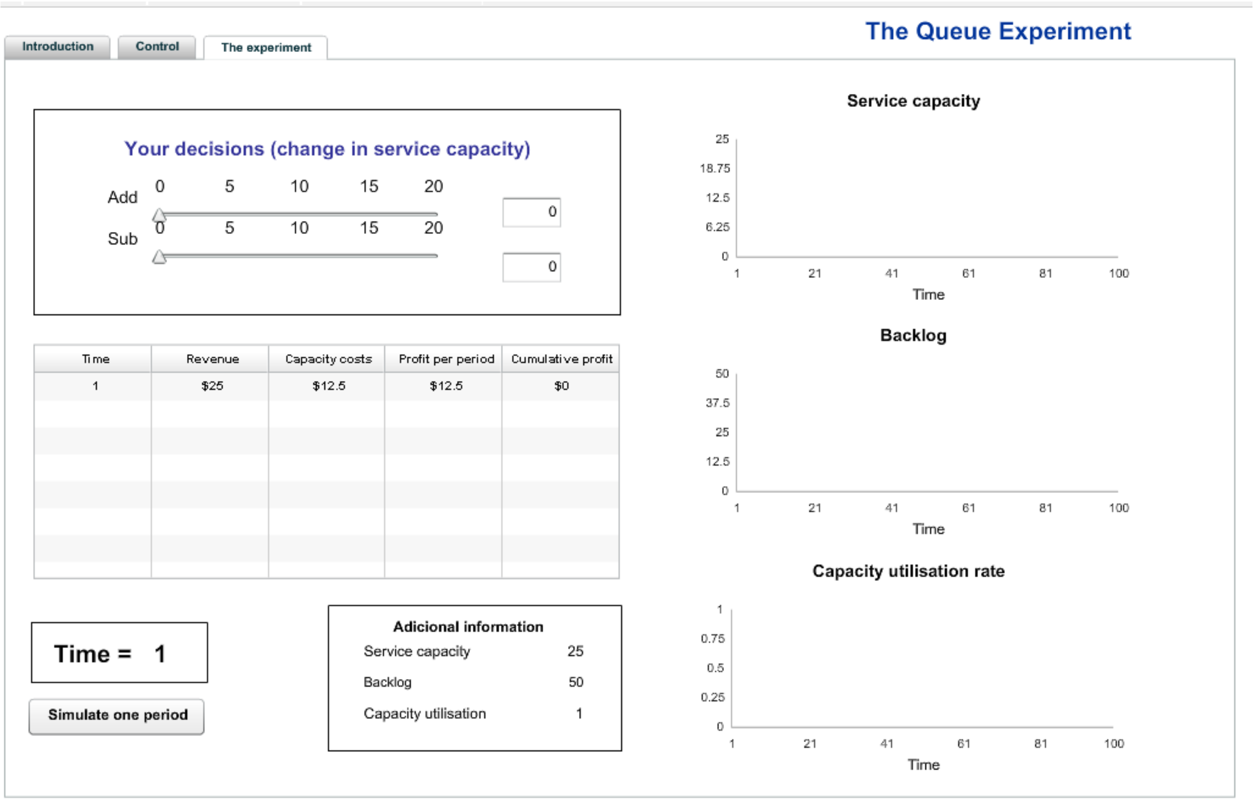
van Ackere, A., Haxholdt, C., & Larsen, E. R. (2006). Long-term and short-term customer reaction: a two-stage queueing approach. *System Dynamics Review*, *22*(4), 349-369.

van Ackere, A., Haxholdt, C., & Larsen, E. R. (2010). Dynamic Capacity Adjustments with Reactive Customers. *Management*. Lausanne.

Yechiali, U. (1971). On optimal balking rules and toll in the GI/M/1 queuing process. *Operations Research*, *19*(2), 349-370.

APPENDIX

A. Computer Interface



Add: 0

Sub: 0

Time	Revenue	Capacity costs	Profit per period	Cumulative profit
1	$25	$12.5	$12.5	$0

Time = 1

Adicional information

Service capacity	25
Backlog	50
Capacity utilisation	1

B. Subjects' Instructions (Base case)

Instructions for the participants

NOTE: PLEASE DO NOT TOUCH THE COMPUTER BEFORE BEING ASKED TO DO SO

Welcome to the experiment on decision making in a service industry. The instructions for this experiment are quite simple. If you follow them carefully and make good decisions, you may earn a certain amount of money. The money will be paid to you, in cash, at the end of the experiment. You are free to halt the experiment at any time without notice. If you do not pursue the experiment until the end, you will not receive any payment. The University of Lausanne has provided funds to support this experiment. If you have any questions before or during the experiment, please raise your hand and someone will come to assist you.

We assure you that the data we collect during the course of this experiment will be held in strict confidence. Anonymity is guaranteed; information will not be reported in any manner or form that allows associating names with individual players.

Description of Experiment

This experiment has been designed to study how managers adjust service capacity in a service facility. Below is a short explanation of the system that you will have to manage in the experiment. It is a relative simple system and you only have to make two decisions each time period (increasing capacity and/or decreasing capacity).

The situation

You are the manager of a large garage, which repairs and maintains cars. You have an existing customer base as well as many potential customers who currently are not using your services, but might consider doing so in the future. Both groups are sensitive to the waiting time.

Waiting time: is the average time between the moment a customer calls your garage to make an appointment and the time the car has been serviced. This depends on two factors, how many other customers have made reservations previously (i.e. how long is the queue) and the service capacity of the garage (i.e. how many cars can on average be serviced per time period). Due to planning constraints, this waiting time cannot be less than one month.

Customers: These customers use your garage on average every twice a year. They evaluate the expected waiting time (which is based on (an average of) the last few times they have used your garage) and compare this expected waiting time to the time they consider acceptable (the average for the industry, which is 2 months: the elapsed time between the moment a customer calls, and the moment he can pick up his car after servicing averages 2 months). If they are satisfied (i.e. the expected waiting time is comparable to or better than the average for the industry) they will remain your customer and return again to use your garage. If they consider that the waiting time is too long compared to the industry average they will switch to another garage.

Potential customers: These are people who might become customers if they consider that your waiting time is attractive (i.e. less than the industry average). However, given that they are currently not among your customers, they only hear about the waiting time at your place through word of mouth. Consequently, their estimate of the waiting time at your place is based on less recent information than the estimate of your current customers. Note: the number of potential customers is unlimited.

Service Capacity: This is the number of cars the garage can service on average in one month. You, as the manager, control the service capacity of the garage, i.e. you have the possibility to increase and/or decrease capacity. However, this cannot be done instantaneously: it takes 4 months to increase capacity (e.g. ordering more tools, hiring people, acquiring more buildings etc) and 2 months to decrease capacity (end a lease on a building, lay off people, etc). Note: If at some point your decisions result in a service capacity equal to zero (0), the garage will be closed and the experiment is ended.

Your Task

As the manager, you make decisions regarding any change in capacity for the garage each month. To help you make these decisions you have information about the number of customers currently waiting for service or whose car is currently being serviced (referred to as the queue), profit, the current capacity of the garage, and the capacity utilization rate. You goal is to maximize the total profit over 100 months.

Cost and revenue information:

Profits [E$/month] = Revenue – Cost

Revenue [E$/month]

= number of customers served [cars/month]*Average Price per Customer [E$/car]

Average Price per Customer = 1 $/car

Cost [E$/month]

= Service capacity [units]* Unit cost of service capacity [E$/unit/month]

Unit cost of service capacity = 0.5 $/unit/month

Interface

In front of the computer, you will have the interface where all interactions will take place. The information is the same as what we have provided in these instructions. Please ask the facilitator to have a trial run to test out the software.

Payment

At the end of the experiment, you will receive a cash reward. This will consist of a guaranteed participation fee of 20CHF, plus a bonus which will depend on the total profit you have achieved. This bonus will vary between 0 and 60CHF. ***If you do not pursue the experiment until the end, you will not receive any payment.***

You will be asked to complete and sign a receipt with your name, email address, and student ID number. Thereafter, you can collect your payment. We will be happy to answer any questions you may have concerning this experiment.

If you want to participate in this experiment, please sign the consent form on your desk. ***This form must be signed before the start of the experiment***

If you have no further questions, please ask the experiment facilitator to begin. Good luck and enjoy the experiment.

Appendix E.

“A queuing system with risk-averse customers: Sensitivity Analysis of Performance.”

Delgado, C. A., A. van Ackere, and E. R. Larsen.

In *Industrial Engineering and Engineering Management (IEEM), 2011 IEEE International Conference on, Singapore, December 2011*. 1720-1724. IEEExplore Digital Library, 2011.

A Queuing System with Risk-Averse Customers: Sensitivity Analysis of Performance

C.A. Delgado[1], A. van Ackere[1], E.R. Larsen[2]
[1]Faculty of Business and Economics, University of Lausanne, Lausanne, Switzerland
[2]Institute of Management, University of Lugano, Lugano, Switzerland
(carlos.delgado@unil.ch, ann.vanackere@unil.ch, erik.larsen@usi.ch)

Abstract - **In this paper, we incorporate decision rules based on adaptive behaviour in order to analyze the impact of customers' decisions on queue formation. We deviate from most of the literature in that we model dynamic queuing systems with deterministic and endogenous arrivals. We apply a one-dimensional cellular automata in order to model the research problem. We describe a self organizing queuing system with local interaction and locally rational customers. They decide which facility to use considering both their expected sojourn time and their uncertainty regarding these expectations. These measures are updated each period applying adaptive expectations and using customers' experience and that of their local neighbours. This paper illustrates how the average sojourn time of customers in the system depends on their characteristics. These characteristics define how risk-averse customers are as well as how conservative they are regarding new information.**

Keywords – **Cellular Automata, Queuing System, Sensitivity Analysis, Simulation, Uncertainty.**

I. INTRODUCTION

Queuing systems may be described as a process where customers arrive at a facility for service. Arrivals are the inputs of the process, while the outputs are served customers [1]. This process includes the service and the wait, when servers are not immediately available. Many researchers in Operation Research, Economics, Management and Computer Science have focused on studying problems related to Queuing systems. Queuing problems have been extensively tackled and discussed since Erlang [2] published his work on telephone traffic problem in 1909. However, most research on this subject has been mainly aimed at the optimization of performance measures and the equilibrium analysis of a queuing system. The early works concerning queuing problems were confined to the equilibrium theory [3] and aimed at design, running and performance of facilities.

Traditionally, analytical modelling and simulation have been the approaches used to deal with queuing problems. The analytical approach describes mathematically the operating characteristics of the system in terms of the performance measures, usually in "steady state" [4]. This method is useful for low-complexity problems whose analytical solutions are not difficult to find. For complex problems, a simulation approach is preferable as it enables modelling the problem in a more realistic way, with fewer simplifying assumptions [4].

The decision process of the customers has been rarely studied. For instance, the way customers as autonomous agents decide which facility to join for service. In a queuing system, customers react to the congestion in queues and adapt their expectations regarding sojourn time using their previous experiences. Koole and Mandelbaum [5] have suggested the incorporation of human factors as a challenge in order to advance the development of queuing models. The seminal papers on this subject are [6] and [7]. Most of the models in this field are stochastic (e.g.[6]; [7]; [8]; [9]; [10]) and their form of feedback is either state-dependent (e.g. [6]) or steady state (e.g. [8]). Some authors have included cost allocation as a control for system congestion (queue size) (e.g. [8]). In this way, customers' decisions on whether or not to join the system are based on such cost. Likewise, those decisions can be based on steady-state analysis (e.g. [8]) or be state-dependent (e.g. [11]). The stochastic models are aimed at understanding the impact of variability of the service and arrival processes on the system behaviour. In spite of all these models and research, little emphasis has been placed on the impact of individual choice on queue formation and understanding of the effects of expectations and experiences. Moreover, the existing models do not consider how customers include uncertainty in their perceptions of sojourn time to decide which facility to join.

By studying how customers make decisions in a queuing system, we attempt to understand in a more realistic way the behaviour of the system in contrast to that research which tries to optimize performance measures using analytical methodologies. Our goal is to build a new basis for the analysis of queuing problems by incorporating decision rules based on adaptive behaviour for customers (e.g. [12]). We ignore exogenous factors, which can influence a customer's decision to seek service, such as: quality of service, added value services and discounts. That is, all facilities work under the same conditions. We deviate from most of the literature in that we model dynamic queuing systems with deterministic and endogenous arrivals. In this way, some assumptions of classical queuing theory are relaxed, in particular that the system reaches a steady-state, that service rates are exogenous and that both service and arrival patterns are stochastic.

Consider a situation where customers routinely require a service and autonomously choose a facility in a multichannel system with one queue for each channel (facility). There are other applications, in which

978-1-4577-0739-1/11/$26.00 ©2011 IEEE

customers decide at what time to patronize a facility for service instead of choosing which facility to join. In these cases, we can consider each period as a service channel. The customers' decision to return in the next period to the same facility, and therefore their loyalty, will depend on their previous experience. Some examples of this kind of systems include: a person who must choose a garage for the inspection of her car, a person who goes monthly to the bank to pay her bills, and a person who goes weekly to the supermarket, among others. In all these examples, the customer may choose the facility at which he wishes to be served and at what time to do so.

We propose an extension of the model of [12] and [13]. We model a queuing system with endogenous and deterministic arrival rates using an agent-based simulation approach, more precisely a one-dimensional cellular automata [14]. We use this model to explain how customers interact in a multichannel service facility and to study their collective behaviour. We describe a self-organized queuing system with local interaction and locally rational customers who, based on their expectations, decide which facility to use. We apply adaptive expectations [15] to model how customers update their expected sojourn time based on their own experience and that of their local neighbours. We also introduce uncertainty into the process of formation of agents' expectations in order to analyse how a risk-averse attitude may affect collective behaviour. In this way, we differ from [12,13] since the agents' decision policy in our model considers both the agents expectations and their uncertainty regarding those expectations. In order to model the agents' uncertainty we use the concept of volatility of forecast errors (e.g. [16,17]).

A sensitivity analysis of the model allows us to identify different customer types and analyse the performance of the system depending on the customer type. Model simulations indicate that customers with a high level of risk-aversion perform well (i.e. low sojourn times) when they give significant weight to new information to update their memory regarding expected sojourn time. Customers with an intermediate level of risk-aversion experience low sojourn times when they are reluctant to update both their expectations of sojourn time and variance. Finally, risk-neutral customers and those with low risk-aversion achieve their best performance when they update their expected sojourn times slowly.

II. METHODOLOGY

We propose Agent-Based Modelling to study how customers routinely choose a facility for service. We model the service facility system as a queuing system with endogenously determined arrivals and exogenous service rates. The way customers decide which facility to patronize each period is based on adaptive expectations [15]. The model used in this paper was developed by [18], who adapted the model from [13] by incorporating uncertainty into the information used by customers to decide which facility to join. While [12] and [13] explain in detail the way customers use adaptive expectations to make decisions, [18] describes how uncertainty is considered.

A one-dimensional cellular automata structure [14] is used to represent the interaction among customers and analyze the way they decide which facility to join each period. The cells represent customers and the states cells can take are the facilities between which customers can choice. We assume a population of n customers and m facilities to set up the system. Each customer has exactly two neighbours, one on each side. Each period (t) customers must choose a facility for service. Customers cannot observe queues before they decide which facility to join. However, they use their last experience and that of their neighbours to update their memory regarding the expected sojourn time at each facility using the following equation:

$$M_{ijt+1} = \alpha M_{ijt} + (1-\alpha)\, W_{jt} \tag{1}$$

where M_{ijt} and M_{ijt+1} are the expected sojourn times of customer i for facility j at time t and $t+1$, respectively, i.e. the past and current expectations of customers. W_{jt} is the new information regarding the sojourn time at facility j, which is the same for all customers patronizing this facility. α is the coefficient of expectations [15]. This new information can be the customer's own experience, or that of his quickest neighbour. For $\alpha = 0$, no weight is given to the past, which implies that the expected sojourn time equals the most recent sojourn time. A value $\alpha = 1$ implies no updating of expectations.

Traditional queuing theory considers systems where the arrival rate is less than the service rate, and calculates average sojourn time in steady state [1]. We assume that the arrival rate may occasionally exceed the service rate, implying that steady state is never reached. Sankaranarayanan, Delgado, van Ackere, and Larsen [13] and [12] suggest (2) to estimate a measure of the average sojourn time (W_{jt}) in a transient state.

$$W_{jt} = \frac{\lambda_{jt}}{\mu^2} + \frac{1}{\mu} \tag{2}$$

where μ represents the service rate, which is assumed as fixed and identical for all facilities and λ_{jt} is the arrival rate at each facility j at time t, which can vary depending on customers' decisions. Equation (2), which is inspired by the behaviour of an M/M/1 system, satisfies the well-known Little's Law [13], but remains well defined when $\rho \geq 1$ (Transient Analysis).

Delgado, van Ackere, Sankaranarayanan, and Larsen [18] differs from [12] and [13] in that [18] assume that customers consider both their expectations and the uncertainty of those expectations to decide which facility to patronize. Uncertainty is estimated using the error involved in the expectations which customers make about the sojourn time at the facilities. However, considering that M_{ijt+1} is updated using exponential smoothing, which assumes a weighted average of the two sources of evidence (M_{ijt} and W_{ijt}), we use the concept of volatility forecasting [16] in order to deal with uncertainty similarly to the expectations, i.e. giving a different weight to the new information and the past. As variance is unobservable, we estimate a smoothed variance, s^2_{ijt+1}, using the equation:

$$s^2_{ijt+1} = \gamma * s^2_{ijt} + (1-\gamma) * (W_{ijt} - M_{ijt-1})^2 \tag{3}$$

where γ is the expectations coefficient [16]. The logic behind the parameter γ is similar to that described above for α. In this case, this parameter refers to the updating process of the variance of customers' expectations instead of their expected sojourn times. In the same way that customers update their expectations of sojourn time, they update their estimate of the variance of the sojourn time for their most recently used facility and for that of their best performing neighbour.

Once the variance of expected sojourn time is known, uncertainty is estimated by the standard deviation s_{ijt}. Customers consider this uncertainty and their expectations of sojourn time in order to form an upper bound for their expected sojourn time. The way customers take into account uncertainty depends on how risk-averse they are. We use R in order to define the customers' profile risk. The higher the value of R, the more risk-averse customers are. In this sense, an upper bound measure for the expected sojourn time, UbM_{ijt}, is given by,

$$UbM_{ijt} = M_{ijt} + R * s_{ijt} \tag{4}$$

This upper bound is the measure on which customers base their decision of which facility to join the next period. Each period customers will patronize the facility with the lowest upper bound, i.e. customers update their state by choosing the queue with the lowest value of UbM_{ijt} (c.f. (1)). In the rare case where two or more facilities are tied for the lowest upper bound of expected sojourn time, customers choose among these facilities, giving first preference to their previously chosen facility and second choice to the one previously used by their best performing neighbour. For more details about the model description and the assumptions made by the authors, refer to [12,13,18]

III. RESULTS

In this section, we present a sensitivity analysis with respect to the risk-aversion parameter (*R*) and the expectation coefficients (α, γ). Fig. 1 illustrates how the average sojourn time of customers in the system varies depending on the value of these parameters. Each graph in Fig. 1 shows the average sojourn time of the system as a function of the expectation coefficient of the variance (γ) and the risk-aversion parameter (*R*), for a given expectation coefficient of expected sojourn time (α). The curves in each graph represent the average sojourn time for different values of the expectation coefficient of uncertainty (γ) depending on the risk level of customers (*R*, horizontal axis).

The system is set up with a population of 120 customers and 3 facilities. Each facility has a service rate of 5 customers per unit of time. Each customer is provided with an initial memory for both the expected sojourn time and the variance of this sojourn time for each facility. The initial expected sojourn times are allocated to customers randomly around the optimal average sojourn time, while the variance is initialized at zero.

The model was tested for different initial expected sojourn time (M_{ij0}) allocated to the customers randomly. This allows us to identify that after 1000 periods the system starts to exhibit a certain stability, i.e. the variance of the sojourn time is less than 10% and the customers' behaviour over time can be described as a collective pattern, which is easily defined and characterized by their decisions. Hence, we have run the model over 1500 periods and the average sojourn time was computed over the last 500 period. The results in Fig. 1 are based on 1000 simulations of the model for each combination of the parameters α, γ, and R. The simulations for each combination of parameters were run using the same random seeds.

Before going in the sensitivity analysis, we define the main customer types in Table I We define conservative customers as those who give little weight to new information (i.e. Their or their neighbour's, most recent experience) to update their memory.

TABLE I
CUSTOMER TYPES DEFINED BY THE PARAMETER VALUES

Definition of customer types	**Parameters**
Risk-neutral	$R = 0$
Risk-averse	$R > 0$
Intermediate risk-aversion level	$R \epsilon$ [0.4, 1.2]
Conservative customers - w.r.t. new information regarding average sojourn time - w.r.t. new information regarding the variability	 $\alpha \geq 0.5$ $\gamma \geq 0.5$
Reactive customers - w.r.t. new information regarding average sojourn time - w.r.t. new information regarding the variability	 $\alpha \leq 0.5$ $\gamma \leq 0.5$

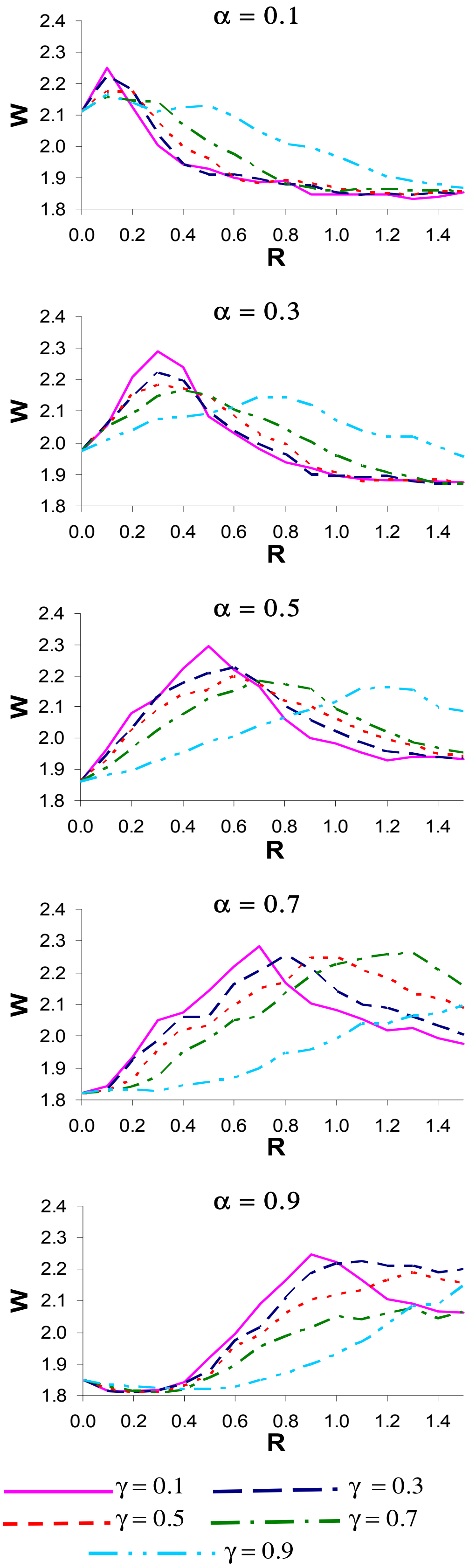


Fig. 1. The average sojourn time of the system as a function of the coefficients of expectations (α, γ) and the risk-aversion parameter (R).

The first graph of Fig. 1 ($\alpha = 0.1$) illustrates the case where customers give more weight to new information than to the past when updating their expectations of sojourn times. In this case, customers with a very low level of risk-aversion (i.e. *R* close to 0) perform poorly, i.e. the system reaches high average sojourn times. On the contrary, a system, where customers are very risk-averse (i.e. *R* large), tends to perform better, i.e. the average sojourn time is lower.

As α increases, very risk-averse customers tend to perform increasingly worse (i.e. higher sojourn times). Whatever the value of α, the maximum sojourn time is always achieved when customers update their perception of the variance more quickly ($\gamma = 0.1$), but the value of *R* (i.e. the risk-aversion parameter) for which this maximum occurs increases in α (i.e. the expectation coefficient to update the expected sojourn time).

When the expected sojourn time coefficient (α) is less than 0.3, very risk-averse customers tend to reach lower sojourn times (i.e. better performance). For values of α greatest than or equal to 0.5 (i.e. more conservative customers), less risk-averse customers perform better. In the extreme case where customers are very conservative regarding their expectations of sojourn time ($\alpha = 0.9$) the average sojourn time achieved by customers with low risk-aversion is close to the Nash Equilibrium, which is equal to 1.8 periods. The system achieves such an equilibrium only when an equal number of customers patronize the three facilities over time. Nash is the optimal behaviour the system could achieve. However, the best performance, which the system achieves, yields a sojourn time a little higher than Nash. This performance occurs when the system is set up with customers characterized as:

- Very conservative to update their expectations of sojourn time (i.e. $\alpha = 0.9$),
- Rather conservative to update their expectations of variance (i.e. $\gamma \geq 0.5$), and
- They have lower risk-aversion level (i.e. *R* є [0.1, 0.3]).

Next, let us consider the impact of the coefficient of expectations to update variance (γ).Customers who are conservative regarding the variance and have an intermediate level of risk-aversion perform well as long as the parameter α remains above 0.3. This is particularly the case when the parameters α and γ are very high (i.e. α and $\gamma = 0.9$, which means that customers are reluctant to consider new information to update their expectations). Very risk-averse customers perform more poorly when they are conservative as regards the variance (γ large) and less conservative regarding their expected sojourn times (α low). The lower γ, the spikier the behaviour of the average sojourn time is as a function of risk-aversion (*R*).

IV. CONCLUSION AND FUTURE WORK

We have applied a one-dimensional cellular automata model to analyze how customers, who patronize a system of service facilities, interact with their neighbours in order to choose the facility with the minimum upper bound of customers' expected sojourn times. Customers compute their upper bound using their expected sojourn time, their estimate of the uncertainty concerning this expectation and their risk-aversion parameter. They estimate their expected sojourn times and level of uncertainty by applying adaptive expectations.

The model has been simulated for different combinations of parameters which characterize customers. These simulations showed that very risk-averse customers experience low sojourn times (i.e. good performance) when they give significant weight to the most recent experience to update their memory regarding expected sojourn time (i.e. α is small). Moreover, they achieve their best performance when they update their perception of the variance more slowly. If these very risk-averse customers are reluctant to take into account new information to update their expectations of sojourn time, they will experience higher sojourn times.

As far as customers with an intermediate risk-aversion level are concerned, they perform better when they update their expectations slowly (both the expected sojourn time and the variance).

Finally, both risk-neutral customers and those with low risk-aversion achieve their best performance when they give little attention to the most recent experience when updating their expected sojourn times.

Future research will include allowing for different levels of reactivity depending on the source of the information, i.e. customers will have different expectation coefficients when updating expectations based on their own experience or that of their neighbours. We will also allow for heterogeneous customers. In particular, we will consider customers with different degrees of risk-aversion (R) and/or different levels of reactivity (α, γ). Another interesting aspect would be to make the service rate endogenous, i.e. the manager would be able to adjust the service capacity of the facilities depending on the customers' behaviour.

ACKNOWLEDGMENT

The authors gratefully acknowledge the support from the Swiss National Science Foundation, grant # 100014_126584/1

REFERENCES

[1] D. Gross and C.M. Harris, Fundamentals of Queueing Theory, New York: Wiley, 1998.

[2] A.K. Erlang, "The theory of probabilities and telephone conversations," Matematisk Tidsskrift, vol. 20, 1909, pp. 33-39.

[3] D.G. Kendall, "Some Problems in the Theory of Queues," Journal of the Royal Statistical Society. Series B, vol. 13, 1951, pp. 151-185.

[4] S.C. Albright and W.L. Winston, Management Science Modeling, South Western Cengage Learning, 2009.

[5] G. Koole and A. Mandelbaum, "Queueing models of call centers : An introduction," Annals of Operations Research, vol. 113, 2002, pp. 41-59.

[6] P. Naor, "The Regulation of Queue Size by Levying Tolls," Econometrica, vol. 37, 1969, pp. 15-24.

[7] U. Yechiali, "On optimal balking rules and toll in the GI/M/1 queuing process," Operations Research, vol. 19, 1971, pp. 349-370.

[8] S. Dewan and H. Mendelson, "User Delay Costs and Internal Pricing for a Service Facility," Management Science, vol. 36, 1990, pp. 1502-1517.

[9] C.M. Rump and S.J. Stidham, "Stability and Chaos in Input Pricing for a Service Facility with Adaptive Customer Response to Congestion," Management Science1, vol. 44, 1998, pp. 246-261.

[10] E. Zohar, A. Mandelbaum, and N. Shimkin, "Adaptive Behavior of Impatient Customers in Tele-Queues : Theory and Empirical Support," Management Science, vol. 48, 2002, pp. 566-583.

[11] A. van Ackere, "Capacity management : Pricing strategy , performance and the role of information," International Journal of Production Economics, vol. 40, 1995, pp. 89-100.

[12] C.A. Delgado Alvarez, A. van Ackere, E.R. Larsen, and K. Sankaranarayanan, "Collective Behavioral Patterns in a Multichannel Service Facilities System : A Cellular Automata Approach," Operations Research, Computing, and Homeland Defense. 12th INFORMS Computing Society Conference, ICS 2011, R.K. Wood and R.F. Dell, eds., Monterey, CA: INFORMS, Hanover, MD, 2011, pp. 16-28.

[13] K. Sankaranarayanan, C.A. Delgado Alvarez, A. van Ackere, and E.R. Larsen, "The Micro-Dynamics of Queuing: Understanding the formation of queues," 2010, pp. 1-24.

[14] M.J. North and C.M. Macal, Managing Business Complexity. Discovering Strategic Solutions with Agent-Based Modeling and Simulation, New York: Oxford University Press, 2007.

[15] M. Nerlove, "Expectations and Cobweb Phenomena," The Quarterly Journal of Economics, vol. 72, 1958, pp. 227-240.

[16] J.W. Taylor, "Volatility forecasting with smooth transition exponential smoothing," International Journal of Forecasting, vol. 20, Mar. 2004, pp. 273 - 286.

[17] J.W. Taylor, "Invited Comments on 'exponential smoothing: The state of the art - Part II' by E. S. Gardner Jr.," International Journal of Forecasting, vol. 22, 2006, pp. 671-672.

[18] C.A. Delgado Alvarez, A. van Ackere, K. Sankaranarayanan, and E.R. Larsen, "Modelling decisions under uncertainty in a queuing system," 25th EUROPEAN Conference on Modelling and Simulation, ECMS 2011, Krakow, Poland: 2011.

Appendix F. List of Symbols

Chapter 2

$\mathcal{A}$: set of customers (i.e. cells) in the CA model

$\mathcal{Q}$: set of facilities (i.e. states) in the CA model

$\mathcal{S}$: set of states $s_i(t)$

e_{ijt}: error in the estimation of M_{ijt}

K: neighbourhood range of the CA model, i.e. how many neighbours each customer has on each side

m: number of facilities at $\mathcal{Q}$

M_{ijt}: expected sojourn time of customer *i* for facility *j* at time *t*

n: number of customers at $\mathcal{A}$

R: risk-aversion factor allocated to the customers

$s_i(t)$: state chosen by customer *i* at time *t*

SS_{ij}: sum of the squared errors

tsim: simulation time

U_{ijt}: expected upper bound of the sojourn time of customers *i* for facility *j* at time *t*

W_{ijt}: average sojourn time experienced by customer *i* at facility *j* at time *t*

$\bar{W}_t$: weighted average sojourn time

α: coefficient of expectations to estimate the expected sojourn time

γ: coefficient of expectations to estimate the smoothed variance

λ_{jt}: arrival rate at facility *j* at time *t*

μ: service capacity of the facilities

σ_{ijt}: uncertainty of customer *I* regarding his expectation of sojourn time for facility *j* at time *t*

σ^2_{ijt+1}: smoothed variance

Chapter 3

d^+: delivery delay

d^-: dismantling delay

M_{ijt}: expected sojourn time of customer *i* for facility *j* at time *t*

W_{ijt}: average sojourn time experienced by customer *i* at facility *j* at time *t*

x_{jt}: capacity orders for facility *j* at time *t*

y_{jt}: capacity retirements from facility *j* at time *t*

α: coefficient of expectations to estimate the expected sojourn time using own information

β: coefficient of expectations to estimate the expected sojourn time using the quickest neighbour's information

δ: coefficient of expectations of managers

$\Delta\mu_{jt}$: required capacity adjustment

ζ: speed at which managers make decisions

θ: coefficient of expectations

λ_{jt}: arrival rate at facility *j* at time *t*

$\hat{\lambda}_{jt}$: future arrival rate at facility *j* at time *t*

$\dot{\mu}_{jt}$: desired service capacity for facility *j* at time *t*

μ_{jt}: available service capacity of facility *j* at time *t*

μ_{jt}^{+}: capacity on order

μ_{jt}^{-}: capacity to be retired

$\mu_{jt}^{\pm}$: capacity decisions not yet implemented

$\bar{\mu}_{jt}$: service capacity which manager *j* perceives to have at time *t*

τ_{MR}: reference average sojourn time

ψ: coherence factor

Chapter 4

ASC_t: available service capacity

CAD_t: capacity adjustment decision

CbR_t: capacity to be retired

CO_t: capacity on order

ΔC_t: capacity decisions not yet implemented

DC_t: desired service capacity

EQ_t: estimating queue (i.e. Perception of the backlog of work)

FSC_t: future service capacity

$PGap_t$: perceived gap between the desired service capacity and the future service capacity managers face each period

Q_t: backlog of customers

RSC_t: retained service capacity

W_t : sojourn time

$\overset{\bullet}{W}_t$: expected sojourn time

δ: coefficient of the manager's expectations (i.e. *1/δ* represents the reference average sojourn time of the manager)

μ_t: service rate

τ_{MR}: market reference

φ: coefficient of customers' expectations (i.e. 1/φ is assumed to be the time taken by customers to adapt their expectations)

χ: speed at which the manager choose to close this perceived gap, i.e. how aggressive is he when making decisions.

ψ: coherence factor