INTERNATIONAL TRANSACTIONS IN OPERATIONAL RESEARCH

WILEY

INTERNATIONAL TRANSACTIONS IN OPERATIONAL RESEARCH

Intl. Trans. in Op. Res. 29 (2022) 2705–2727
DOI: 10.1111/itor.12836

Service facilities with risk-averse customers: a simulation approach

Carlos A. Delgado-Alvarez[a,*], Ann van Ackere[b] and Erik R. Larsen[c]

[a] *Faculty of Business, Politécnico Colombiano Jaime Isaza Cadavid, Carrera 48 No. 7-151, Medellin, Colombia*
[b] *Faculty of Business and Economics (HEC), University of Lausanne, Quartier UNIL-Dorigny Bâtiment Anthropole, Lausanne 1015, Switzerland*
[c] *Department of Management, Aarhus Universitet Business and Social Sciences, Fuglesangs Allé 4, Aarhus V DK-8210, Denmark*
E-mail: cadelgado@elpoli.edu.co [Delgado-Alvarez]; ann.vanackere@unil.ch [van Ackere]; erik.larsen@mgmt.au.dk [Larsen]

Received 28 November 2019; received in revised form 21 May 2020; accepted 21 May 2020

Abstract

In this study, a one-dimensional cellular automata model is used to represent a self-organized queueing system with local interaction between captive and boundedly rational customers who repeatedly choose a facility for service. While previous work has focused on decision rules based on adaptive expectations, the present work expands this analysis by explicitly incorporating customers' attitude toward risk to study the impact of risk aversion on the collective behavior and the average system sojourn time. The customers' decision process is modeled using adaptive expectations and incorporating the uncertainty involved in these expectations. Customers update their expectations based on their own experience and that of their neighbors. Simulation analysis is used to compare the aggregated behavior for different degrees of customer risk aversion. Risk-neutral customers base their decisions only on their expected sojourn time, while risk-averse customers account for uncertainty by estimating an upper bound of the sojourn times. The results indicate that the more risk-averse the customers, the longer the transient period, and the more slowly the system converges to an almost stable average sojourn time. Systems where customers have an intermediate level of risk aversion achieve the worst average sojourn times.

Keywords: decision-making under uncertainty; simulation; service operations management; adaptive expectations; sojourn times

1. Introduction

Understanding people's attitudes to risk, and the resulting impact on their behavior, has been of interest to researchers from different areas over the last decades. This includes the pioneering work

*Corresponding author.

Published by John Wiley & Sons Ltd, 9600 Garsington Road, Oxford OX4 2DQ, UK and 350 Main St, Malden, MA02148, USA.

in the field of decision science (Kahneman and Tversky, 1979; Tversky and Wakker, 1995), and more recently in finance (Keller and Siegrist, 2006; Fellner and Maciejovsky, 2007), project management (Yang et al., 2016), supply chains (Guo and Liu, 2020), and climate change (Leiserowitz, 2006; Whitmarsh, 2008), among others.

Risk attitudes affect the way in which people take decisions when facing choices relating to queueing; this influences the formation of queues and the resulting sojourn times. A well-known phenomenon concerns the impact of commuter behavior on the timing of rush hours: risk-averse commuters leave increasingly early in an attempt to beat the traffic, resulting in rush traffic starting earlier. Similarly, travelers who fear travel delays, arrive increasingly early at the airport, which leads to overcrowded facilities. Other examples relate to the choice of facility. For instance, a customer faces the choice between the local post office, where the waiting time is highly variable (often no queue at all, occasionally a very long one) and the main post office, where the queue length is more predictable as service capacity is flexible. A similar dilemma concerns traveling to a supermarket, where one is sure to find the desired product, versus going to the local corner store, with a higher probability of not finding what one wants and having to look elsewhere. In a traveling context, one can opt for the main road, facing a long but fairly predictable travel time, versus opting for smaller roads, which may be faster on average but may face extremely long delays in case of an incident.

Risk-averse customers are likely to base their choice not on the expected or most likely sojourn time, but on a worst-case estimation. This results in asymmetric reactions to extreme events: while a (very) bad experience will significantly increase the worst-case expectation, a (very) good experience has a much more moderate impact. This issue is the focus of the present paper. The model considers repeat customers who must routinely choose a facility for service and who interact with their neighbors sharing information about their experience. They use their previous experience and that of their best-performing neighbor to update their expectations regarding sojourn times. The "neighbors" could for instance represent colleagues at work discussing their experience driving in that day. Customers are aware that these expectations may not be accurate and consider this uncertainty to a certain degree, depending on how risk-averse they are. More specifically, they estimate an upper bound for their expected sojourn time and choose which facility to join based on this upper bound. The impact of updating expectations based on information from all neighbors rather than only the best-performing neighbor is elaborated upon in the discussion.

Hassin (2016, p. 1) classifies queueing systems research into three categories: performance analysis, optimal design and control, and analysis of rational strategic behavior. Traditional research has given more attention to the first two topics; more recently, behavioral aspects have received increasing attention when studying queueing problems. This paper pertains to the last category: it focusses on investigating how customers' risk attitude influences their decisions when choosing a facility, and how this affects the resulting collective behavior and the average sojourn time of the system.

The remainder of the paper is structured as follows. After a literature review, the system being studied is described. Section 4 introduces the model and provides a technical description. Next, the concept of adaptive expectations is introduced, with an explanation of how it is used to estimate the expected sojourn times and the customers' uncertainty regarding these expectations. Section 5 characterizes the customer types according to their level of risk aversion and the weights they give to their expectations. This is followed by an explanation of the simulation results for a typical case and a sensitivity analysis with respect to the risk-aversion parameter and the expectation coefficients.

Next, we consider what happens when a customer deviates from the behavioral rules assumed in the model. The last section addresses the conclusions and future work.

2. Literature review

While waiting is generally considered to negatively affect quality, there are exceptions. Kremer and Debo (2015) study customer behavior in a queueing situation where long waits are synonymous with good quality, that is, customers consider facilities with large queues a good place to patronize. They provide theoretical and experimental evidence that the probability of new customers (uninformed customers) joining a facility for service decreases when waiting times are short—"the empty restaurant syndrome." Still, in this work the focus is on the more common case where customers dislike waiting.

Depending on the information decision makers have available when making their choice, behavioral queueing models can be classified as models with unobservable queues, with partially observed queues, or with observable queues (Hassin, 2016, p. 8). Unobservable queue models consider that real-time information on the system state is not available for decision makers. Observable queue models consider that decision makers have full information about the system state (i.e., the queue length and the state of the server). In between these two extremes, partially observed queue models assume that decision makers have partial information about the system state (e.g., the queue length or the state of the server). Altman (2005) presents another classification of behavioral queueing models according to the questions that arise for customers: to queue or not to queue; when to queue; and where to queue. The model proposed in this paper is a behavioral queueing model with unobservable queues and customers deciding where to queue.

Most research on behavioral queueing concentrates on studying equilibrium behavior and social optimality. This stream of literature started when Naor (1969) published his seminal papers in which he formalizes and quantifies the insights published by Leeman (1964). Naor (1969) studies an $M/M/1$ queueing model with an observable queue where homogenous customers decide whether to join the queue depending on an admission fee (i.e., a toll). Edelson and Hildebrand (1975) extend Naor's model to unobservable queues. Hassin and Haviv (2003) provide a review of the literature building on Naor's work until 2003, while Hassin (2016) covers the more recent literature.

The marketing literature also contributes to this area by studying the influence of waiting times on customer satisfaction, customer loyalty, and service quality (Law et al., 2004; Bielen and Demoulin, 2007). This line of work was highlighted by Koole and Mandelbaum (2002) who emphasize the need to include human factors in queueing models as a challenge to advance the development of the area.

In more recent work, Hassin and Snitkovsky (2017) study equilibrium strategies in a queueing problem with unobservable queues, where customers must choose between waiting in line to receive a free service or pay to be served immediately. Roet-Green and Hassin (2017) analyze the impact of information acquisition on the decision of whether to join or balk. Pala and Zhuang (2018) study equilibrium strategies in a security screening process using an $M/M/1/K$ queueing system with generalized process sharing, with heterogeneous impatient customers who balk once the waiting time exceeds their patience time. Hassin and Koshman (2017) study Naor's model: they consider different contexts (e.g., limited information, constraints on price flexibility) and derive easily implementable mechanisms enabling achieving close to optimal profits. Wang et al. (2018) focus on systems characterized by highly seasonal arrival rates, for example, food courts facing a lunch-hour rush, using a deterministic fluid model. They develop a rule of thumb for capacity decisions and consider both monopoly and duopoly situations.

The present work differs from this stream of literature in that it analyzes the resulting collective behavior of customers according to their risk attitudes rather than developing equilibrium strategies or identifying optimal capacity or pricing strategies. The model assumes captive customers who routinely must choose a facility for service during rush time and are not allowed to balk or jockey; queues are unobservable before a decision is made.

Risk attitudes, although highly relevant, have received little attention in this context. Previous research shows that when customers make decisions involving waiting time, they typically tend to be rather risk-averse (Leclerc et al., 1995). Consequently, they prefer the service provider with more predictable service times. From this point of view, service providers must focus on reducing the uncertainty of their service times in order to make their facilities more attractive. Kumar and Krishnamurthy (2008) show that when customers account for uncertainty on both supply and demand sides, the uncertainty about the expected congestion eliminates this risk aversion, unless customers expect to face congestion everywhere or nowhere. They argue that customers' decision strategies in queueing situations are complex and that using mean-variance utility models, risk aversion (i.e., choosing the facility with the lower uncertainty in the service time), or crowd avoidance (i.e., avoiding joining the facility where one expects the majority to go) is not sufficient to understand the complexity behind customers' behavior at service facilities. More recently, Wang and Zhang (2018) perform a detailed equilibrium analysis of an $M/M/1$ queueing system, assuming strategic risk-sensitive customers, allowing for a continuum from risk-seeking to risk-averse customers, from three points of view: individual self-interest, profitability of the facility, and social welfare. They consider different information contexts (e.g., observable vs. unobservable queue) and are able to show that several standard queueing theoretic results derived under the assumption of risk neutrality no longer hold with risk-sensitive customers.

Most of the literature reviewed above does not consider feedback, that is, customers' current behavior is not affected by information about past periods. One of the reasons is that this leads to nonlinear models, which mostly do not have closed-form solutions. Agnew (1976) is one of the first to explicitly incorporate feedback: he provides stationary solutions for a system where service capacity varies with queue length. Haxholdt et al. (2003) go a step further: they consider a deterministic model where both arrival rate and service rate are endogenous, depending, respectively, on customers' perception of past waiting times and on queue length. Modeling the arrival rate as a function of past waiting times is the first step toward explicitly modeling repeat customers, as done by Law et al. (2004) and van Ackere et al. (2013).

Behavioral models that enable studying the link between the individuals' decisions and aggregated performance include van Ackere and Larsen (2004) who analyze how commuters form expectations about congestion in a three-road system using a one-dimensional cellular automata (CA) model. Sankaranarayanan et al. (2014) use a similar approach to model a multichannel service facility where rational customers choose one facility for service based on their expected sojourn time, which they compute using their own experience and that of their neighbors. Then, Delgado-Alvarez et al. (2020) elaborate upon Sankaranarayanan et al.'s (2014) model by enabling managers to adjust the facilities' service capacity. The present work extends Sankaranarayanan et al. (2014) in several directions, with the aim of capturing more realistic individual behavioral patterns, while making the link to more traditional OR models: (a) the concept of volatility of forecast errors (Taylor, 2004, 2006) is used to incorporate uncertainty into the process of expectation formation and (b) this enables an analysis of how different levels of risk aversion affect the collective behavior of customers.

The simulation results indicate that the more risk-averse customers are, the longer the transient period exhibited by the system is, and the longer the system takes to converge to an almost stable average sojourn time. Systems where customers exhibit an intermediate level of risk aversion perform worse and are more likely to end up ignoring a facility (i.e., never use it again) than those whose customers are risk-neutral or strongly risk-averse.

3. A service system

An agent-based simulation model, structured as a one-dimensional CA (Wolfram, 1994, p. 411), is developed to represent a queueing system with endogenous and deterministic arrival rates. It is used to explain how customers, who must repeatedly choose a facility for service, interact in a multichannel system; the focus is on how their collective behavior is influenced by their level of risk aversion. Examples of such systems include choosing a store for weekly shopping, a garage for annual car maintenance and commuters.

The model represents a self-organizing queueing system with local interaction between captive rational customers. Consider a fixed population of n homogenous customers. All customers need service simultaneously and must choose one of m facilities. This system portrays a captive market in which customers repeatedly need either a service or a good and have several options to obtain it. The assumption that customers need service simultaneously is a stylized representation of a rush hour, where queues form at the different facilities as 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 real-time information regarding the queue size or sojourn time when making their choice and cannot leave or switch a facility after making their choice.

Before deciding, agents are assumed to communicate with their neighbors and share information about their previous experience. In each period, customers use their most recent experience and that of their best-performing neighbor to update their expectations of the sojourn time at the different facilities. They do so using an adaptive expectations process (Nerlove, 1958) to combine their memory and their new information (i.e., their most recent experience and that of their quickest neighbor). Customers who give more weight to their memory than to new information are labeled "conservative." In contrast, those giving more weight to new information are called "reactive." Uncertainty is incorporated into the agents' expectation formation process in order to analyze how a risk-averse attitude may affect collective behavior. The agents' uncertainty is quantified using the concept of volatility of forecast errors (Taylor, 2004). Risk-neutral customers choose which facility to use based on their expected sojourn times (i.e., they ignore uncertainty). Risk-averse customers base their decision on their estimate of the upper bound of the expected sojourn times, which they compute using their expected sojourn times and their estimate of the uncertainty, taking into account their risk-aversion level.

Occasionally, a customer may deviate from the bounded-rationality assumptions outlined above, whether by mistake or by choice (i.e., using a different decision rule, see, for instance

Delgado-Alvarez et al., 2017). To the outsider such behavior may appear random. We therefore consider a variation of the model that incorporates this possibility.

4. The cellular automata model

The system described above is modeled as a queueing system with endogenous arrival rates and exogenous service rates. While the facilities have identical service rates, their arrival rates depend on the customers' previous experiences, their expectations, and their risk aversion. Reneging, balking, and jockeying are not allowed.

A CA approach is adopted to represent the interaction between customers, model their expectations, and analyze their collective behavior (Wolfram, 1994, p. 412; North and Macal, 2007, pp. 46–49). In the context of the CA methodology, agents portray customers. A one-dimensional ring-shaped neighborhood is assumed, where each cell is an agent who interacts with exactly one neighbor on each side: agent i communicates with agents $i-1$ and $i+1$. As the cells are located on a ring, all cells have a neighbor on each side. Without this ring structure, the two agents located at the extremes of the CA would each have only one neighbor; this would require defining tailor-made decision rules for these two agents, referred to as boundary conditions, complicating the model without yielding additional insights. The neighborhood can depict, for example, a social network consisting of neighbors, colleagues, members of a sports club, etc.

In each period, agents use a simple decision rule, based on their memory of past experience and new information, to choose a facility (state). The updating process of agents' memory is based on the theory of adaptive expectations (Nerlove, 1958), also known as exponential smoothing: agents weight their most recent information and their memory when forming expectations (Theil and Wage, 1964). The information stems from their own experience and that of their neighbor who achieved the lowest 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 using the squared error of the forecasts. Exponential smoothing is used to estimate the squared residuals (Taylor, 2006), a method known as volatility of forecasting errors (Taylor, 2004).

To formalize the CA model, A denotes the set of n agents (cells) $\{A_1, A_2,\ldots, A_i,\ldots, A_n\}$ and Q the set of m possible facilities (states) $\{Q_1, Q_2,\ldots, Q_j,\ldots, Q_m\}$. 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}), which depend on customers' decision at time t. The arrival rate is a function of the state of the agents each period, $s_i(t)$. The agents' decisions will determine their state in each period. Let S denote 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, $S \subset \{Q_1, Q_2,\ldots, Q_j,\ldots, Q_m\}$. 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:

$$x_{ij}(t) = f(s_i, Q_j, t) = \begin{cases} 1 & \text{if } s_i(t) = Q_j \\ 0 & \text{otherwise} \end{cases} \quad \forall\, i = 1, 2, \ldots, n \quad j = 1, 2, \ldots, m \tag{1}$$

$$\lambda_{jt} = \sum_{i=1}^{n} x_{ij}(t) \quad \forall\, j = 1, 2, \ldots, m. \tag{2}$$

The state $s_i(t)$ evolves over time depending on the agent's expected sojourn times, his uncertainty and his risk aversion. The expected sojourn time of agent A_i for facility Q_j in period t is denoted by M_{ijt}, the corresponding uncertainty by σ_{ijt}, and the risk-aversion factor by R, where a larger value of R denotes a higher degree of risk aversion. 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 $s_i(t)$ of agent A_i is determined by the following function:

$$s_i(t) = F(M_{ijt}, \sigma_{ijt}, R) \quad \forall\, i = 1, 2, \ldots, n \quad j = 1, 2, \ldots, m. \tag{3}$$

In order to define this function, these variables are incorporated into a single measure that enables agents to decide their state each period, labeled the *upper bound of the expected sojourn time* and denoted by B_{ijt}, which represents the maximum sojourn time that agents estimate they might experience at facility j. This upper bound is estimated using the agents' expected sojourn time (M_{ijt}) and uncertainty measure, and their risk-aversion level. Agents choose to patronize the facility with the lowest upper bound. Before delving into the function that determines the upper bound of the expected sojourn time (B_{ijt}), a description of how agents form their expectations and estimate their uncertainty by applying an adaptive expectation model is provided. A table summarizing the notation of the model is given in the Appendix (see Table A1).

4.1. The adaptive expectations model

Agents update their expected sojourn time M_{ijt} and their uncertainty σ_{ijt} by applying adaptive expectations, a forecasting method commonly applied to financial and economic time series (Gardner, 2006). In each period, a new estimate is obtained as the weighted average of the most recent observation and the previous estimate. Agents update their expected sojourn time and their uncertainty for the facility they have selected, and for the facility selected by their quickest neighbor. When these two facilities coincide, agents only update their memory and their uncertainty for this facility, using their own information. For this updating process, they use Equations (4)–(6), as explained in the following sections.

4.1.1. 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 neighbors ($i - 1$ and $i + 1$) 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 neighbor. With this information in mind, agents update their expected sojourn time for their chosen facility and that of their best-performing neighbor using an exponentially weighted average with weight α, which is assumed to be constant. Given the assumption of homogenous agents, α is the same for all agents. So, the updating process of the agents' memory (M_{ijt+1}) can be expressed as

$$M_{ijt+1} = \alpha M_{ijt} + (1-\alpha)W_{ijt} \quad \forall\, i = 1, 2, \ldots, n, \tag{4}$$

where M_{ijt} is the previous value of the memory, and j refers to the last facility used by the agent and that used by his best-performing neighbor.

4.1.2. Estimating the uncertainty of the expected sojourn times

Using the concept of volatility forecasting (Taylor, 2006), the estimation error e_{ijt}, that is, the difference between the previously estimated value M_{ijt-1} and the most recently observed value W_{ijt}, is calculated:

$$e_{ijt} = W_{ijt} - M_{ijt-1} \quad \forall\, i = 1, 2, \ldots, n. \tag{5}$$

Next, exponential smoothing is applied to the squared estimation errors (Taylor, 2004): the smoothed variance, σ^2_{ijt+1}, is calculated as the weighted average of the previous estimate, σ^2_{ijt}, and the new observation of the squared error e^2_{ijt}. Thus, agents update the variance, σ^2_{ijt}, as follows:

$$\sigma^2_{ijt+1} = \gamma\sigma^2_{ijt} + (1-\gamma)e^2_{ijt} \quad \forall\, i = 1, 2, \ldots, n, \tag{6}$$

where γ denotes the smoothing constant. Taking the square root yields the measure of volatility σ_{ijt}. Again, this updating process is only applied to the facility (j) used by the agent and that used by his best-performing neighbor.

4.1.3. Estimating the upper bound of the expected sojourn time

Once the agents have computed their expected sojourn time, and its uncertainty, they consider these values to assess the upper bound of the expected sojourn time, that is, the estimate of the maximum sojourn time they think they could experience given their expectations, their uncertainty regarding these expectations, and their risk aversion. Given the aforementioned "risk-aversion factor," R, which is identical for all agents, the upper bound of the expected sojourn time, B_{ijt}, of agent A_i at facility Q_j in period t can be written as follows:

$$B_{ijt} = M_{ijt} + R\sigma_{ijt} \quad \forall\, i = 1, 2, \ldots, n \quad j = 1, 2, \ldots, m. \tag{7}$$

The second term of the right-hand side of the equation can be interpreted as a safety margin: the higher an agent's uncertainty concerning his estimate of the average sojourn time (σ_{ijt}) and the higher his risk aversion (R), the larger this time buffer. Agents will patronize the facility with the lowest value of B_{ijt}, that is, the lowest upper bound of the expected sojourn time. Should two or three facilities be tied for the lowest expected upper bound, a very rare occurrence, agents will choose among these, with their current facility as the first choice, and the one of their best-performing neighbor as second choice.

4.2. Average sojourn time in a transient period

This study focuses on a system in transient state, where the arrival rates can temporarily exceed the service rates. Therefore, the expected sojourn time measure proposed by Sankaranarayanan et al. (2014) is used:

$$W_{jt} = \lambda_{jt}/\mu^2 + 1/\mu \quad \forall\ j = 1, 2, \ldots, m, \tag{8}$$

where μ denotes the service rate and λ_{jt} the arrival rate, that is, the number of agents arriving at facility j in period t. This measure remains well defined for $\rho \geq 1$ (transient analysis), while satisfying Little's law and the steady-state equations (Gross et al., 2008, pp. 9–10).

5. Characterization of customers

Customers (referred to as agents in this paper according to the terminology of CA models) can be classified according to the values of their behavioral parameters. The different customer types are characterized as follows:

- Depending on the coefficients of expectations (α and γ), customers can be called reactive or conservative regarding new information. For values of $\alpha < 0.5$, customers are considered reactive regarding the expected sojourn times, since they attach more importance to the new information than to the past. Alternatively, if $\alpha > 0.5$, customers are considered conservative with regard to the expected sojourn times, that is, they give little weight to the new information regarding their or their neighbors' most recent experience. A similar reasoning applies to γ: Customers are considered to be either reactive or conservative regarding the use of new information to estimate the variability of their expectations when, respectively, $\gamma < 0.5$ or $\gamma > 0.5$.
- Customers are considered risk-neutral when $R = 0$, that is, they ignore the uncertainty. If $R > 0$, they account for uncertainty. They are referred to as having a low, intermediate, or high level of risk aversion depending on the value of R. In the discussion, the values of 0.4 and 1.2 are used as cutoff points between these three groups.

6. Simulation results and discussion

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) and a neighborhood size equal to 1. The service rate is the same for all facilities and equals five agents per unit of time. Each agent is provided with an initial memory for the expected sojourn time for each facility. These memories are distributed randomly around the optimal average sojourn time. The system behavior depends on the initial values of memory assigned to the agents, that is, the evolution of the system is path dependent. Although customers' initial memory is allocated randomly, behavior generated by the model is not stochastic: the waiting times are deterministic functions of the number of

Table 1
Parameter values used for the simulation run

Parameter	Value	Description
m	3	Number of service facilities
n	120	Population size
μ	5	Service rate
α	0.4	Weight to memory when updating the expected sojourn time
γ	0.6	Weight to memory when updating the estimated variance of the expected sojourn time
R	0, 0.5, 1.5	Risk-aversion parameter: risk-neutral, intermediate risk aversion, high risk aversion, respectively
$Tsim$	100	Simulation time for the illustrative example
	1500	Simulation time for the in-depth analysis
Seed	1	Random seed used to initialize the model (agents' initial memory)

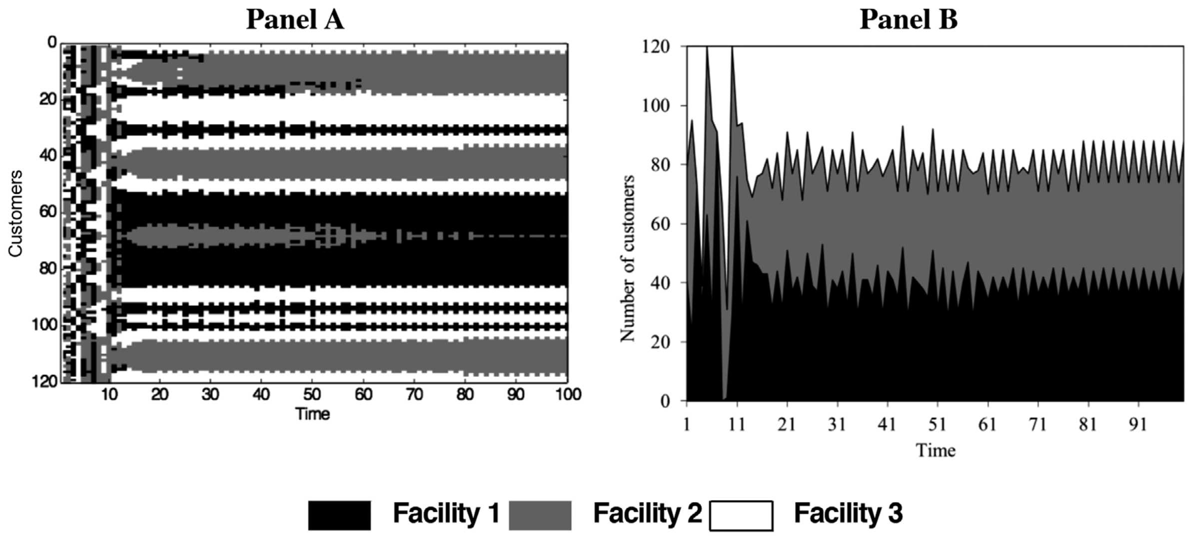

Fig. 1. Spatial–temporal behavioral evolution of risk-neutral agents' choice of service facility (Panel A) and stacked chart of the number of customers patronizing each facility over time (Panel B) for $\alpha = 0.4$; $\gamma = 0.6$; and $R = 0$.

customers joining a facility (Equation (8)), and customers' decisions are deterministic functions of their information (Equations (4)–(7)). The simulation model is implemented in MATLAB (2009).

6.1. Base case

Table 1 summarizes the parameters used to configure the system for the base case. The chosen time horizons are 100 periods for the illustrative example given below to provide an intuitive understanding of the model behavior, and 1500 for the in-depth analysis.

Figures 1 and 2 compare the evolution of the agents' decisions for different risk attitudes: risk-neutral ($R = 0$, Fig. 1) and moderately risk-averse ($R = 0.5$, Fig. 2) over 100 periods (horizontal

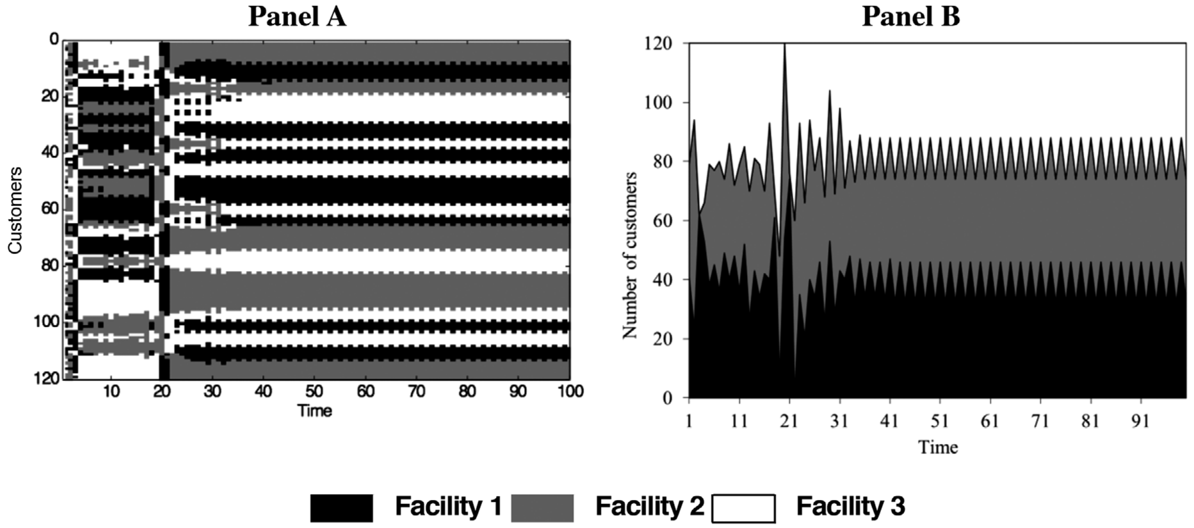

Fig. 2. Spatial–temporal behavioral evolution of risk-averse agents' choice of service facility (Panel A) and stacked chart of the number of customers patronizing each facility over time (Panel B) for $\alpha = 0.4$; $\gamma = 0.6$; and $R = 0.5$.

axis). Panel A shows the spatial–temporal behavior, while the stacked graph in panel B illustrates the evolution of the total number of customers at each facility over time. To ensure comparability (same initial values of expected sojourn times), the same random seed (a value of 1) is used. Each line represents the sequence of choices of an agent (black, gray, and white for, respectively, facilities 1, 2, and 3).

The behavior observed in Figs. 1 and 2 is explained in two phases: the first phase is a period of exploration where agents gather information, learn about the system, and imitate their best-performing neighbor; this is referred to as the transition period. During this period, customers' behavior creates a phenomenon of herding: it is not uncommon for one or two facilities to be very crowded (e.g., periods 7–10 in Fig. 1, periods 18–22 in Fig. 2). Consequently, customers update their memory based on these poor experiences, and move away from heavily used facilities. For instance, in Fig. 1, after being overcrowded at time 7, facility 1 is empty at time 8 and one customer (customer 94) returns in period 9. Figure 2 exhibits a similar phenomenon over the periods 19–21.

After the transient period the system stabilizes, and a pattern of collective behavior emerges. In certain cases, agents experiencing a short sojourn time communicate with their neighbors who in turn join this facility. For instance, in Fig. 1, agents 10 and 11 have a good experience (i.e., low sojourn time) at facility 2 in period 11, and over time attract back a cluster of about 15 customers who remain at that facility. In other cases, a group of neighbors simultaneously moves back to a facility, has a positive experience, and stays there as they are no longer exposed to new information. An example of this can be seen in Fig. 2, where agents 81–96 move to facility 2 in period 22, and most stay there until the end of the simulation. Such groups of agents, who stay at a facility, are referred to as loyal. They are separated by "switchers": agents who keep moving between two facilities according to a regular pattern. The vast majority of switchers alternate between two facilities. For instance, in Fig. 1, agent 3 alternates between facilities 2 and 3. Switchers can exhibit more

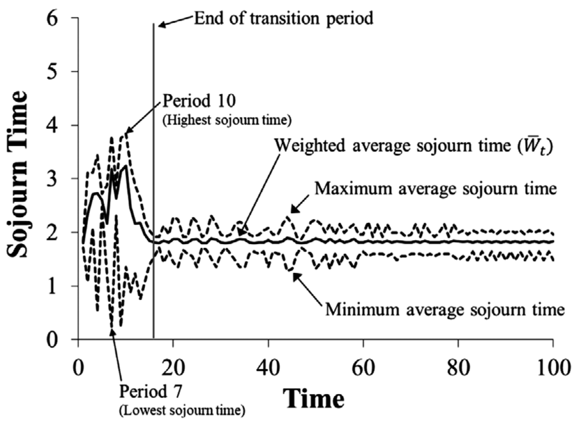

Fig. 3. Average, maximum, and minimum sojourn time for a system configured with risk-neutral agents ($R = 0$) and coefficients of expectations: $\alpha = 0.4$; $\gamma = 0.6$.

complex patterns, for instance agent 68 from period 84 onward in Fig. 1: 1-2-1-2-2-2-1-2-1-2-2-2-, etc.

Panel B of Figs. 1 and 2 provides a clearer picture of how the distribution of customers across the three facilities changes over time: during the exploration period we observe strong fluctuations, with occasionally one facility not being patronized at all (e.g., facility 2 (gray) at time 3 and facility 3 (white) at times 5 and 10 in Fig. 1) and another one being heavily patronized (e.g., facility 2 at time 10 in Fig. 1 and time 20 in Fig. 2). Once the system stabilizes, the allocation of customers follows a regular pattern (from time 79 onward in Fig. 1 and time 35 in Fig. 2). Had a Nash equilibrium been reached, the allocation of agents across the three facilities would have remained constant. When agents update their expectations based on the information from both neighbors, the qualitative behavior is similar, but the system takes much longer to stabilize.

Figures 3 and 4 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:

$$\bar{W}_t = \sum_{j=1}^{3} W_{jt}\lambda_{jt}/120, \tag{9}$$

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

As mentioned above, during the transition agents tend to overpatronize certain facilities in a given period; this results in extreme values for the minimum, maximum, and average sojourn times. For instance, in the risk-neutral case (Figs. 1 and 3), most agents choose facility 1 in period 7, experiencing a very high sojourn time ($W_{1,7} = 3.80$), while at the same time the few agents at facility 2 are much better off ($W_{2,7} = 0.24$). The risk-averse agents (Figs. 2 and 4) have a similar experience in period 21: the two agents at facility 2 experience the lowest sojourn time ($W_{2,21} = 0.28$), while the many agents at facility 1 have a bad day (($W_{1,21} = 3.16$).

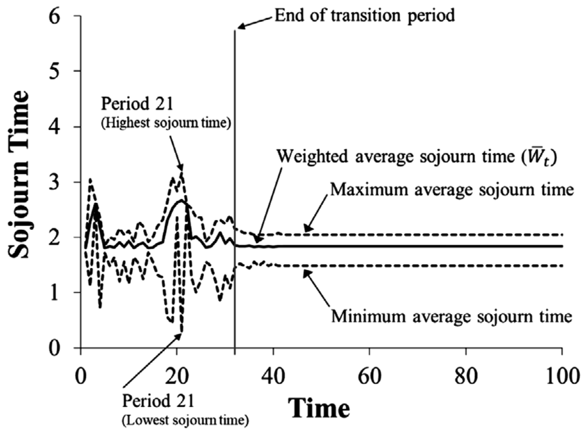

Fig. 4. Average, maximum, and minimum sojourn time for a system configured with risk-averse agents ($R = 0.5$) and coefficients of expectations: $\alpha = 0.4$; $\gamma = 0.6$.

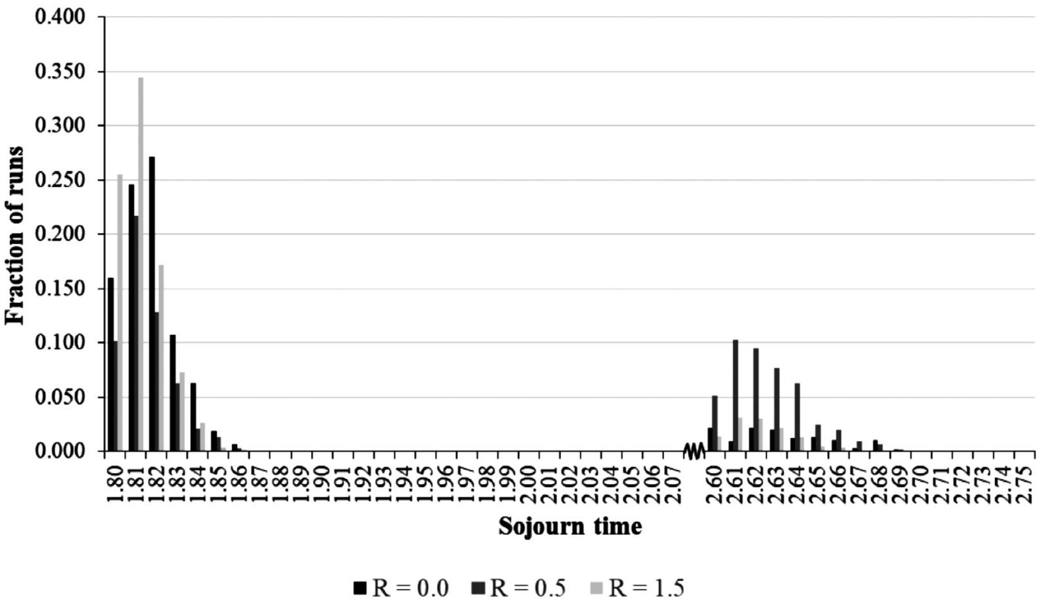

Fig. 5. Distribution of the steady-state weighted average sojourn time for 1000 simulations with different initial conditions for $\alpha = 0.4$; $\gamma = 0.6$ for risk-neutral ($R = 0$), moderately risk-averse ($R = 0.5$) and highly risk-averse ($R = 1.5$) agents.

Note. We use a bin size of 0.01. As the model does not yield observations in the range [2.0800, 2.5999], a "break" indicator is inserted to shorten the axis and improve legibility of the graphs.

After these illustrative examples, a more in-depth analysis of the steady-state system behavior is provided for three levels of risk aversion: indifference to risk ($R = 0$, black bins in Fig. 5), an intermediate degree of risk aversion ($R = 0.5$, dark gray bins in Fig. 5), and a very high degree of risk aversion ($R = 1.5$, light gray bins in Fig. 5). Using the same parameters (see Table 1), the model is run 1000 times for each level of risk aversion. Extensive experimenting with different initial conditions has shown that the transition period rarely exceeds 1000 periods: by that time, a stable pattern of collective behavior has emerged and the variance of the sojourn time is less than 10%. Hence, a runtime of 1500 periods has been selected and the weighted average sojourn time is calculated over the last 500 periods of each run. Comparatively, when agents update their expectations based on the information from both neighbors (as opposed to only the best-performing neighbor), the system takes up to 10,000 periods to stabilize for small values of R, and even longer for larger values of R.

Figure 5 illustrates how the steady-state sojourn time varies depending on the initial conditions. The distribution is bimodal, whatever the level of risk aversion. This is a consequence of the possibility that one of the facilities is not patronized in steady state, that is, it was driven out of business due to a long-term lack of customers. The first peak of each distribution corresponds to the runs where the three facilities remain in use. This scenario accounts for over 80% of the runs when the agents are risk-neutral or highly risk-averse, compared to barely more than 50% for an intermediate level of risk aversion ($R = 0.5$). When all facilities remain in use, the weighted average sojourn time in steady state ranges between 1.80 and 1.95, with many values close to 1.80. The second peak represents the cases where one of the facilities has closed down in steady state, with most values clustered around 2.62. Note the broken line between these two regions, indicating the truncated horizontal axis, which highlights the gap between the two peaks (no values between 1.95 and 2.60).

Based on this analysis, one can conclude that the more risk-averse the customers, the longer the transient period the system exhibits. Moreover, Fig. 5 shows that there is a nonmonotonic relationship between the degree of risk aversion and system performance: customers with an intermediate degree of risk aversion are more likely to cluster in two facilities, thereby driving the third one out of business. Consequently, very risk averse customers and risk-neutral customers achieve lower average sojourn times. Both conclusions remain valid when agents update their expectations using information from both neighbors.

6.2. Sensitivity analysis

A detailed sensitivity analysis is performed with respect to the three main behavioral parameters (the risk-aversion parameter (R) and the expectation coefficients (α, γ)) using the definitions of customer types given in Section 4. The other settings are unchanged (see Table 1). Figure 6 illustrates how the weighted average sojourn time varies as a function of these parameters. Each panel in Fig. 6 corresponds to a specific value of the expectation coefficient of the variance (γ). In each panel the different curves represent the steady-state weighted average sojourn time as a function of the risk-aversion parameter R (horizontal axis, steps of 0.1), for different values of the expectation coefficient of the expected sojourn time (α). For each parameter combination 1000 iterations were performed, using the same 1000 random seeds. The run length is again set at 1500 periods and the weighted average sojourn time is computed for the last 500 periods.

As mentioned above, the weighted average sojourn time always exceeds 1.8 and values falling in the interval [2.6, 2.8] indicate that a facility has closed down. Thus, the higher the weighted average sojourn times exhibited in Fig. 6, the higher the probability that in steady state only two facilities remain in operation for the corresponding parameter combination.

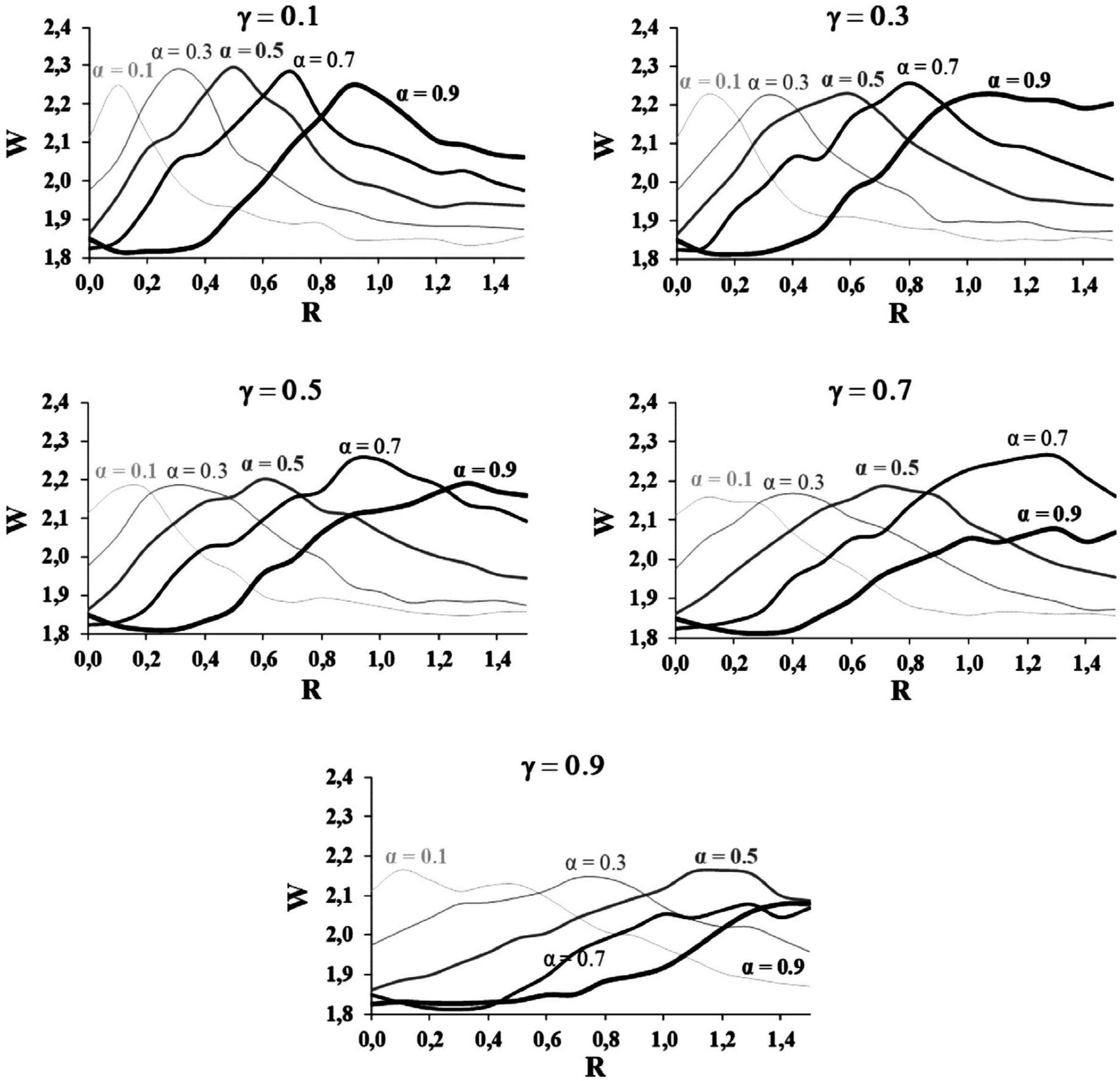

Fig. 6. The weighted average sojourn time of the system as a function of the coefficients of expectations (α, γ) and the risk-aversion parameter (R).

A facility j is forced to close down when all customers' estimated upper bound of the sojourn time (B_{ijt}) at this facility is well above that of the other two facilities. When this occurs, customers will not receive new information to update B_{ijt} over the next periods. Consequently, they will either stay at one of the two remaining facilities or alternate between these.

The first panel of Fig. 6 ($\gamma = 0.1$) illustrates the case where customers give significantly less weight to the past than to new information when updating their expectations of the variance of the

Table 2
Overview of relationship between risk-aversion and updating parameters

Attitude toward new information regarding expected sojourn time (α)	Level of risk aversion, parameter R			
	Risk-neutral	Low	Intermediate	High
Reactive (low α)	–	–	+, if $\gamma \approx \alpha$	+
Intermediate	+		–	
Conservative (high α)	+	+, if γ high	+, if $\gamma \approx \alpha$	–

Note: (–) Combinations yielding high sojourn times, whatever the value of γ; (+) combinations yielding desirable sojourn times (for an appropriate choice of γ).

sojourn times. For a given value of α (i.e., the expectation coefficient to update the expected sojourn time), the performance is nonmonotonic in the level of risk aversion, with the worst performance occurring when α and R are of the same magnitude. For low levels of risk aversion, performance improves in α, while for high levels of risk aversion, the reverse is true. Note that when R exceeds 1 (the maximum value which α can take), the weighted average sojourn time decreases in R for most cases. Looking at the other panels allows us to conclude that these observations are quite robust to the value of γ.

Figure 6 provides the following insights into the performance of customers depending of the weight they give to their memory: for $\alpha \leq 0.3$ (i.e., reactive customers according to the customer types of Section 5), very risk averse customers ($R > 1.2$) perform better than risk-neutral customers ($R = 0$); while for $0.5 \leq \alpha \leq 0.7$ (i.e., rather conservative customers), risk-neutral customers do better than risk-averse customers ($R > 0$).

Whatever is the value of γ, the value of R for which it is most likely that a facility closes down (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 perform the worst. On the contrary, when customers are very conservative ($\alpha = 0.9$) with regard to their expected sojourn time and very reactive with regard to the variance ($\gamma = 0.1$), the average sojourn time starts decreasing in R once this parameter exceeds a certain threshold.

The lowest sojourn time the system can achieve is the Nash equilibrium (1.8 periods), when agents remain distributed evenly across the three facilities. Values close to this performance occur most frequently when customers with a low level of risk aversion ($0.1 \leq R \leq 0.3$) have a very conservative attitude toward updating sojourn time expectations ($\alpha = 0.9$) and are comparatively less conservative when updating expectations of variance ($\gamma < 0.9$).

Next, consider the impact of γ, the coefficient of expectations to update the estimate of the variance. The higher γ, the less sensitive the system behavior is to small changes in the level of risk aversion (R). Indeed, looking at the different panels, the curves become less spiky as γ increases. The performance of truly conservative customers (high γ and α) deteriorates as their risk aversion increases. Very risk averse customers perform better when they consider all the new information as very important (low γ and α).

Table 2 provides a summary of the discussion of Fig. 6. Recall that the parameter γ plays no role for risk-neutral customers, as they base their decision only on the expected sojourn time. These customers perform well as long as they are not overly reactive to new information regarding the

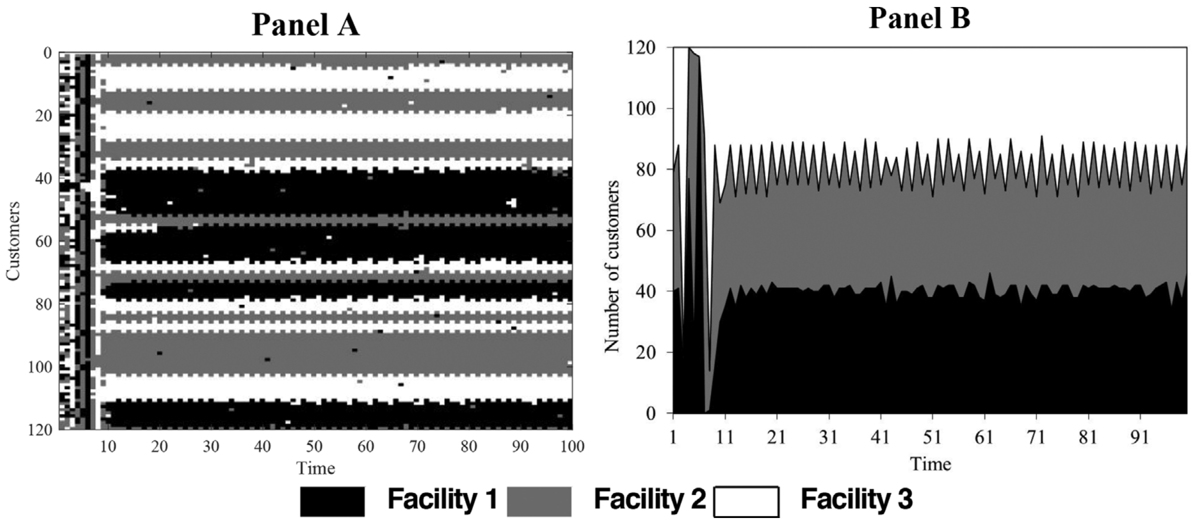

Fig. 7. Spatial–temporal behavioral evolution of risk-averse agents' choice of service facility (Panel A) and stacked chart of the number of customers patronizing each facility over time (Panel B) for the modified model where one randomly chosen customer chooses a facility at random for $\alpha = 0.4$; $\gamma = 0.6$; and $R = 0.0$.

sojourn time. Somewhat risk-averse customers perform best if they are conservative toward new information (high α and γ), while highly risk-averse customers should be reactive when updating their expectation of the average sojourn time. Customers with an intermediate level of risk aversion should have the same attitude when updating their sojourn time and variability expectations: either high values for α and γ, or low values for both.

One would expect highly risk-averse customers to exhibit a conservative attitude toward new information (high α and γ). The results indicate that this is the worst possible choice for them. Similarly, one might expect customers who are risk-neutral or have a low risk aversion, to be reactive toward new information. Again, this would lead them to incur the longest average sojourn times.

7. Impact of nonrationality on facility closure

In this section, we consider a modified model where a customer can deviate from the boundedly rational decision rule assumed so far. More specifically, we model a situation where, in each period, one randomly chosen customer chooses a facility at random. This behavior might represent a situation where a customer makes a mistake, has a biased perception of past experience, or deliberately chooses to make a nonrational decision.

To illustrate the impact of this change, we first provide the spatial–temporal behavior and customer distribution for the two illustrative examples discussed in Section 6.1. Figures 7 and 8, respectively, correspond to the examples of Figs. 1 and 2. The top panels illustrate that the microlevel behavior, that is, the spatial–temporal behavior, may or may not be significantly affected by customers occasionally choosing at random. The length of the transition period changes, and this period can be characterized by more or less variability of sojourn time. The magnitude of these differences is

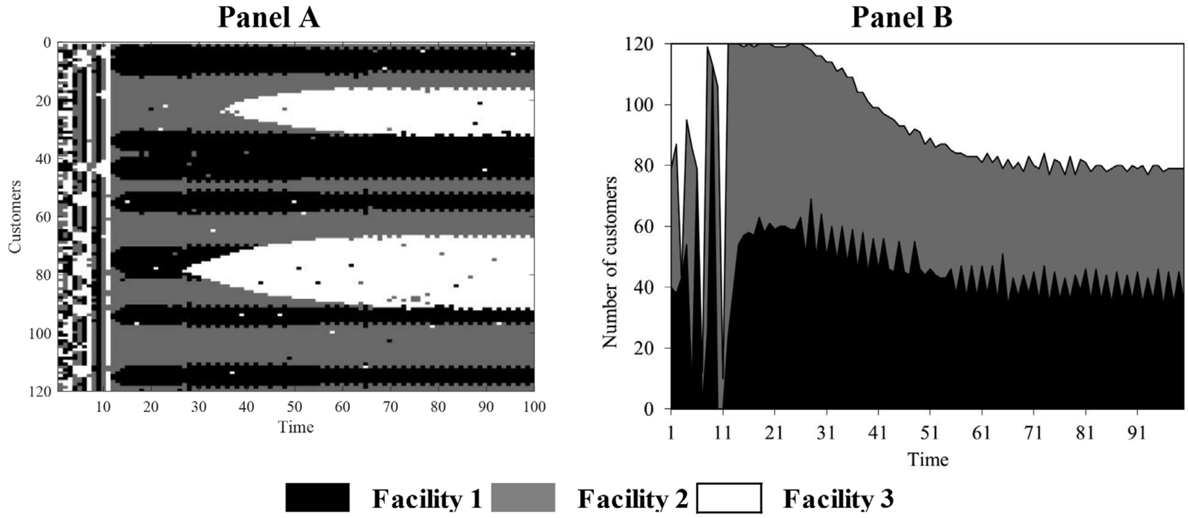

Fig. 8. Spatial–temporal behavioral evolution of risk-averse agents' choice of service facility (Panel A) and stacked chart of the number of customers patronizing each facility over time (Panel B) for the modified model where one randomly chosen customer chooses a facility at random for $\alpha = 0.4$; $\gamma = 0.6$; and $R = 0.5$.

not materially different from that observed between runs of the same model with different random seeds. In other words, both models exhibit similar behavior after the transition period. Panel B of Figs. 7 and 8 illustrates the evolution of how customers allocate themselves between the facilities. The equilibrium behavior is very similar to that observed in Figs. 1 and 2.

The behavior resulting from the two models is quite different if we consider an example where, with boundedly rational agents, one facility closes. Figure 9 illustrates such a case; the left panels show the behavior in the original model (i.e., boundedly rational agents) and the right panels in the modified model (i.e., in each period one randomly chosen customer chooses a facility at random). As expected, the presence of this random element eliminates the closure of a facility. Specifically, if according to the boundedly rational rule a facility should not be patronized again, in the modified model there is, in any given period, one chance in three that the random customer will choose that facility; consequently the probability of no one returning to that facility for the next 10 periods is less than 2%, and well below 1 in a 1000 for the next 20 periods.

Figure 10 illustrates the impact of the initial conditions on the steady-state sojourn time for this modified model. First, as expected, we no longer observe values in the 2.6–2.7 range, which corresponds to the case with only two facilities operating. Next, note that the presence of a random customer implies that in the long run the Nash value of 1.80 will never be achieved, that is, we do not observe situations where in the long run customers remain equally allocated among the three facilities. More generally, the resulting long-term allocation is on average further away from the Nash equilibrium, with the mode of the average sojourn times being around 1.88 (compared to 1.82). This larger spread is particularly striking for $R = 1.5$ (highly risk-averse customers), where over half the values exceed 1.87 (the largest value observed in the base case when the three facilities remain operational, recall Fig. 5).

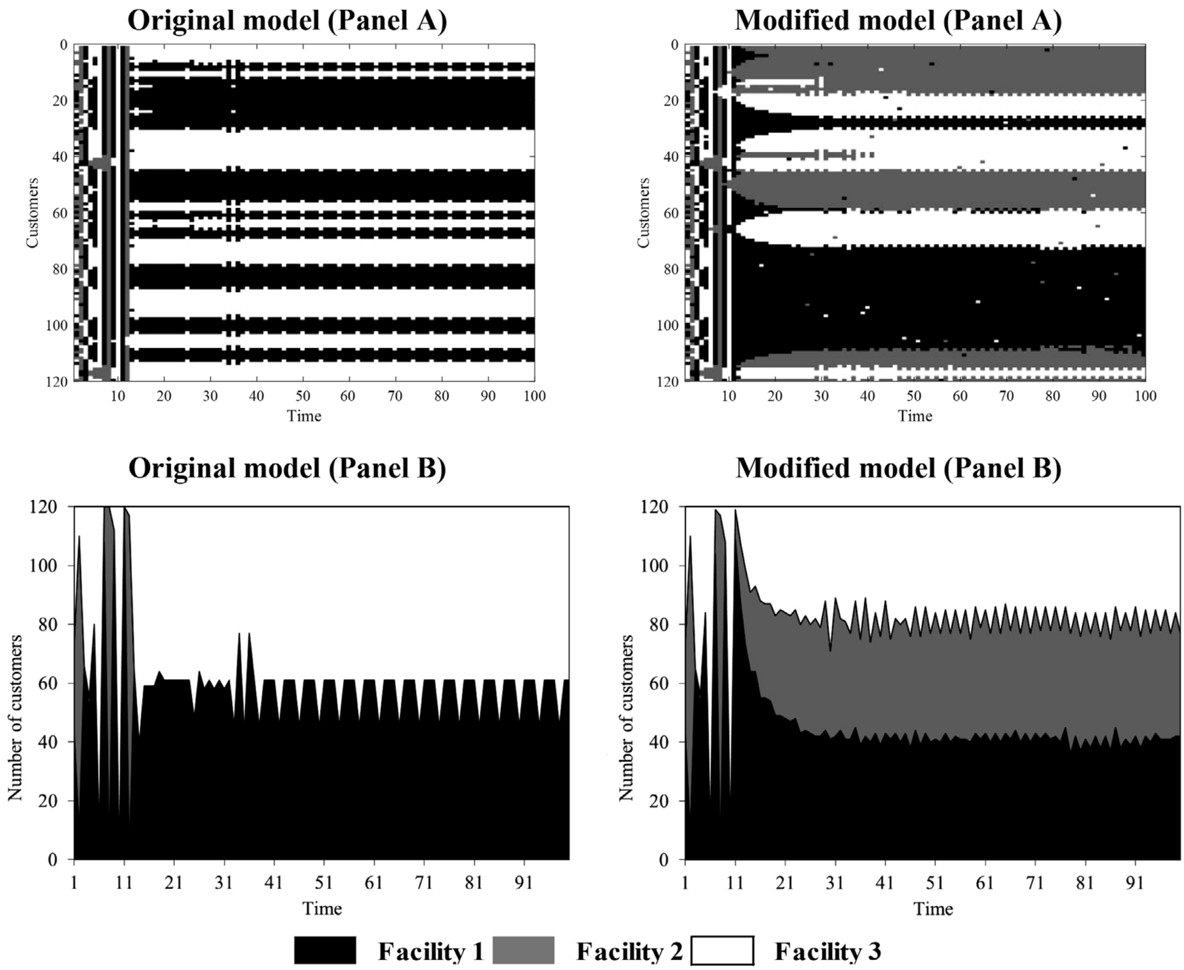

Fig. 9. Spatial–temporal behavioral evolution of risk-averse agents' choice of service facility (Panel A) and stacked chart of the number of customers patronizing each facility over time (Panel B), comparing the behavior in the original model (left) and the modified model (right), for example, where one facility closes in the original model for $\alpha = 0.4$; $\gamma = 0.6$; and $R = 0.5$, seed 3.

8. Conclusions and future work

This paper presents a model of a service system with interacting customers who must decide in each period which facility to join for service. It provides an analysis of the impact of customers' level of risk aversion on the collective behavior 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 neighbors and share information regarding their experiences. They incorporate their information using adaptive expectations. Risk-neutral customers ignore uncertainty; they choose a facility based on their expected sojourn times. Risk-averse customers explicitly incorporate uncertainty in their decision process; they decide based on an estimated upper bound of the sojourn times. This

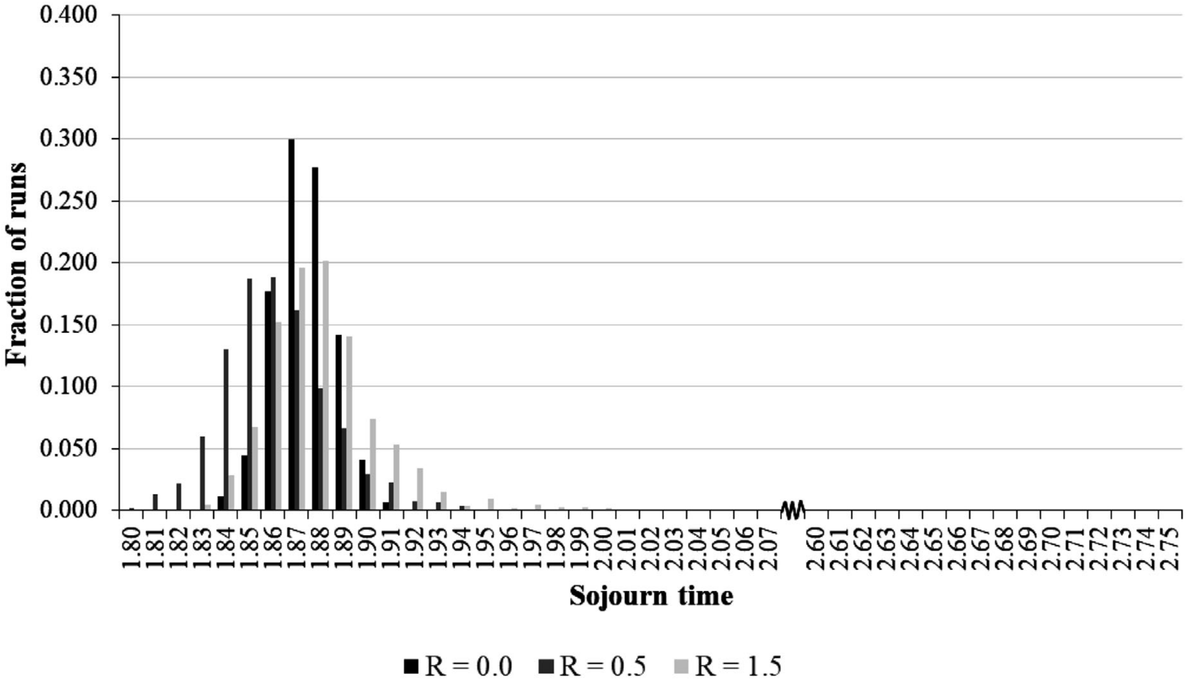

Fig. 10. Distribution of the steady-state weighted average sojourn times for 1000 simulations of the modified model (i.e., one randomly chosen customer chooses a facility at random) with different initial conditions for $\alpha = 0.4$; $\gamma = 0.6$ for risk-neutral ($R = 0$), moderately risk-averse ($R = 0.5$) and highly risk-averse ($R = 1.5$) agents.
Note. We use a bin size of 0.01. Although there are no observed values beyond 2.08, the horizontal axis is identical to that of Fig. 5 for the reason of comparability.

upper bound is estimated based on their expected sojourn times, their estimated level of uncertainty and their degree of risk aversion. We have furthermore considered a stochastic variation of this deterministic model, to analyze the impact of a customer occasionally deviating from the assumed boundedly rational behavior. This eliminates the closure of a facility due to it being deserted by the customers.

Considering the academic implications of this work, the first thing to note is that systems where customers have an intermediate level of risk aversion yield the worst performance. This corroborates the findings of other studies that have shown that in many situations where a decision needs to be made, the worst case is to be caught in the middle (De Toni et al., 2012). In this instance, people should either choose to ignore risk (i.e., base their decision on expected values), or explicitly take it into account in a significant way. A second observation is that highly risk-averse customers perform well if they update their memory of expected sojourn time fast when receiving new information (low α). On the contrary, customers with little or no risk aversion are better off when updating their expected sojourn times cautiously (α large). This implies that people who base their decisions only on expected values (risk-neutral) should treat new information cautiously, while individuals who explicitly incorporate uncertainty in their decision-making should be reactive to new information and incorporate it fast into their decision-making.

Another observation is that, when customers are risk-averse or risk-neutral, it is less likely that one of the facilities is driven out of business than when customers have an intermediate degree of risk aversion. This further emphasizes the previous point, that one should avoid being caught in the middle, as this increases the likelihood of a suboptimal outcome. Finally, customers with an intermediate risk-aversion level should update their expectations regarding average sojourn time and uncertainty in a coordinated way, that is, they should use either two low weights or two high weights.

Turning to the more practical implications of this study, a caveat is in order: one should be cautious about drawing general implications from a stylized model. For instance, in the real world, permanently ignoring a facility is unrealistic; facilities that fail to attract sufficient customers end up closing down: coffee shops, retailers, gas stations, etc., regularly do go out of business. If a facility remains in operation, sooner or later a customer will return, whether by accident or by curiosity (he has not been there for a long time) and experience a very low sojourn time, thereby inducing others to return. To illustrate this behavior, we incorporated an element of random choice in the model, which prevents a facility being overlooked for a long time. As expected, in instances where the three facilities remained operational in the deterministic model, adding this element did not affect the long-term macrolevel behavior.

The implications discussed above allow outlining some heuristics based on the observed behavior. Looking at the combination of risk attitude and the speed at which new information is incorporated into the decision, one observes a form of compensation mechanism. The risk attitude is balanced using new information, that is, risk aversion needs to be offset by a more proactive use of new information, otherwise decisions may end up trailing the current situation, and *vice versa*. However, risk-averse decision makers often tend to be more cautious about incorporating new information, waiting for it to be confirmed over time. Similarly, aggressive decision makers, who are willing to take risks, are likely to use the most recent information, when it would be in their interest to be somewhat more cautions.

The results discussed above point to a link between the customers' risk attitude and the optimal value of the updating parameters, indicating a need for further research on the impact of these behavioral parameters (α, γ, R). This should include allowing for different levels of customer reactivity depending on the source of the information, that is, giving different weights to own experience and information received from neighbors in the memory updating processes. The next step will be to consider heterogeneous customers, in particular, 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 behavior when the facilities have different service capacity or, more interestingly, assuming that managers are able to adjust the service capacity to the number of customers (i.e., endogenous service rates). It would furthermore be interesting to test some of these results in laboratory experiments, as has been done with other models; see, for instance, Delgado-Alvarez et al. (2017).

Acknowledgments

This work was supported by the Swiss National Science Foundation (grants 100012-116564 and 100014-126584).

References

Agnew, C.E., 1976. Dynamic modeling and control of congestion-prone systems. *Operations Research* 24, 3, 400–419.

Altman, E., 2005. Applications of dynamic games in queues. In Nowak, A.S., Szajowski, K. (eds) *Advances in Dynamic Games. Annals of the International Society of Dynamic Games*, Vol. 7. Birkhäuser, Boston, MA, pp. 309–342.

Bielen, F., Demoulin, N., 2007. Waiting time influence on the satisfaction-loyalty relationship in services. *Managing Service Quality* 17, 2, 174–193.

De Toni, A.F., Biotto, M., Ioan, L., 2012. The Pacorini case study: deliberate and emergent strategy. In De Toni, A.F., Franco, R.D., Li, J., Li, Y., Nassimbeni, G., Sartor, M., Zhao, X., Xu, X. (eds) *International Operations Management: Lessons in Global Business*. Gower Publishing, Farnham, UK, pp. 7–32.

Delgado-Alvarez, C.A., van Ackere, A., Larsen, E.R., 2020. Managing service facilities with endogenous arrival and service rates. *International Transactions in Operational Research* 27, 2, 1133–1161.

Delgado-Alvarez, C.A., van Ackere, A., Larsen, E.R., Arango-Aramburo, S., 2017. Managing capacity at a service facility: an experimental approach. *European Journal of Operational Research* 259, 1, 216–228.

Edelson, N.M., Hildebrand, D.K., 1975. Congestion tolls for Poisson queuing processes. *Econometrica* 43, 1, 81–92.

Fellner, G., Maciejovsky, B., 2007. Risk attitude and market behavior: evidence from experimental asset markets. *Journal of Economic Psychology* 28, 3, 338–350.

Gardner, E.S. Jr., 2006. Exponential smoothing: the state of the art—part II. *International Journal of Forecasting* 22, 4, 637–666.

Gross, D., Shortle, J.F., Thompson, J.M., Harris, C.M., 2008. *Fundamentals of Queueing Theory* (4th edn). John Wiley & Sons, Hoboken, NJ.

Guo, S., Liu, N., 2020. Influences of supply chain finance on the mass customization program: risk attitudes and cash flow shortage. *International Transactions in Operational Research* 27, 5, 2396–2421.

Hassin, R., 2016. *Rational Queueing*. CRC Press, Boca Raton, FL.

Hassin, R., Haviv, M., 2003. *To Queue or Not to Queue: Equilibrium Behaviour in Queueing Systems*. Kluwer Academic Publichers, Boston, MA.

Hassin, R., Koshman, A., 2017. Profit maximization in the M/M/1 queue. *Operations Research Letter* 45, 5, 436–441.

Hassin, R., Snitkovsky, R.I., 2017. Strategic customer behavior in a queueing system with a loss subsystem. *Queueing Systems* 86, 3–4, 361–387.

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.

Kahneman, D., Tversky, A. 1979. Prospect theory: an analysis of decision under risk. *Econometrica* 47, 2, 263–291.

Keller, C., Siegrist, M., 2006. Investing in stocks: the influence of financial risk attitude and values-related money and stock market attitudes. *Journal of Economic Psychology* 27, 2, 285–303.

Koole, G., Mandelbaum, A., 2002. Queueing models of call centers: an introduction. *Annals of Operations Research* 113, 41–59.

Kremer, M., Debo, L., 2015. Inferring quality from wait time. *Management Science* 62, 10, 3023–3038.

Kumar, P., Krishnamurthy, P., 2008. The impact of service-time uncertainty and anticipated congestion on customers' waiting-time decisions. *The Journal of Service Research* 10, 3, 282–292.

Law, A.K.Y., Hui, Y.V., Zhao, X., 2004. Modeling repurchase frequency and customer satisfaction for fast food outlets. *International Journal of Quality & Reliability Management* 21, 5, 545–563.

Leclerc, F., Schmitt, B.H., Dube, L., 1995. Waiting time decision making: is time like money? *Journal of Consumer Research* 22, 1, 110–119.

Leeman, W.A., 1964. The reduction of queues through the use of price. *Operations Research* 12, 5, 783–785.

Leiserowitz, A., 2006. Climate change risk perception and policy preferences: the role of affect, imagery, and values. *Climatic Change* 77, 1–2, 45–72.

Naor, P., 1969. The regulation of queue size by levying tolls. *Econometrica* 37, 1, 15–24.

Nerlove, M., 1958. Adaptive expectations and cobweb phenomena. *The Quarterly Journal of Economics* 72, 2, 227–240.

North, M.J., Macal, C.M., 2007. *Managing Business Complexity: Discovering Strategic Solutions with Agent-Based Modeling and Simulation* (1st edn). Oxford University Press, New York, NY.

Pala, A., Zhuang, J., 2018. Security screening queues with impatient applicants: a new model with a case study. *European Journal of Operational Research* 265, 3, 919–930.

Roet-Green, R., Hassin, R., 2017. The impact of inspection cost on equilibrium, revenue, and social welfare in a single-server queue. *Operations Research* 65, 3, 804–820.

Sankaranarayanan, K., Delgado, C., van Ackere, A., Larsen, E.R., 2014. The micro-dynamics of queueing: understanding the formation of queues. *Journal of Simulation* 8, 4, 304–313.

Taylor, J.W., 2004. Volatility forecasting with smooth transition exponential smoothing. *International Journal of Forecasting* 20, 2, 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, 4, 671–672.

The MathWorks Team, 2009. MATLAB 7.9. The MathWorks, Cambridge.

Theil, H., Wage, S., 1964. Some observations on adaptive forecasting. *Management Science* 10, 2, 198–206.

Tversky, A., Wakker, P., 1995. Risk attitudes and decision weights. *Econometrica* 63, 6, 1255–1280.

van Ackere, A., Haxholdt, C., Larsen, E.R., 2013. Dynamic capacity adjustments with reactive customers. *Omega* 41, 4, 689–705.

van Ackere, A., Larsen, E.R., 2004. Self-organising behaviour in the presence of negative externalities: a conceptual model of commuter choice. *European Journal of Operational Research* 157, 2, 501–513.

Wang, H., Olsen, T.L., Liu, G., 2018. Service capacity competition with peak arrivals and delay sensitive customers. *Omega* 77, 80–95.

Wang, J., Zhang, Z.G., 2018. Strategic joining in an M/M/1 queue with risk-sensitive customers. *Journal of the Operational Research Society* 69, 8, 1197–1214.

Whitmarsh, L. 2008. Are flood victims more concerned about climate change than other people? The role of direct experience in risk perception and behavioural response. *Journal of Risk Research* 11, 3, 351–374.

Wolfram, S. 1994. *Cellular Automata and Complexity: Collected Papers* (1st edn). Westview Press, Boulder, CO.

Yang, K., Zhao, R., Lan, Y., 2016. Impacts of uncertain project duration and asymmetric risk sensitivity information in project management. *International Transactions in Operational Research* 23, 4, 749–774.

Appendix

Table A1
Notation

Parameter	Description
μ	Service rate
m	Number of service facilities
n	Population size
R	Risk-aversion factor
α	Weight to memory when updating the expected sojourn time
γ	Weight to memory when updating the estimated variance of the expected sojourn time
Variable	**Description**
λ_{jt}	Arrival rate for queue j at time t
M_{ijt}	Expected sojourn time of agent i for facility j at time t
W_{ijt}	Sojourn time experienced by agent i at facility j at time t
W_{jt}	Sojourn time at facility j at time t
$\bar{W}_t$	Weighted average sojourn time at time t
σ^2_{ijt}	Estimated variance of the expected sojourn time of agent i for facility j at time t
B_{ijt}	Upper bound of the expected sojourn time of agent i at facility j at time t
e_{ijt}	Estimation error of the expected sojourn time of agent i for the facility j at time t