



IRM

INSTITUTE OF RESEARCH IN MANAGEMENT
INSTITUT DE RECHERCHE EN MANAGEMENT

Dynamic Capacity Adjustments with Reactive Customers

WORKING PAPER 0814

Ann van Ackere,

Christian Haxholdt,

and

Erik R Larsen

Dynamic Capacity Adjustments with Reactive Customers

Ann van Ackere
HEC Lausanne
Internef
Université de Lausanne
CH 1015 Dorigny-Lausanne
Switzerland
Tel: +41 21 692 3454
e-mail: ann.vanackere@hec.unil.ch

Christian Haxholdt
Center for Statistics
Copenhagen Business School
Solbjerg Plads 3
DK-2000 Frederiksberg
Denmark
Tel: +45 3815 3513
e-mail: ch.mes@cbs.dk

Erik R Larsen
Institute of Management
University of Lugano
Via G. Buffi 13
CH-6904 Lugano
Switzerland
Tel: +41 58 666 3649
e-mail: erik.larsen@lu.unisi.ch

December 2008

Working paper

Please do not cite or quote

Dynamic Capacity Adjustments with Reactive Customers

Abstract

In this paper we develop a behavioural model in which customers come and go based on their perception of waiting time (relative to other facilities) while managers gradually adjust the capacity of the facility based on their perception of demand. We explicitly account for the difference in access to information between existing and potential customers, which implies that the perception of potential customers lags the perception of current customers. We investigate the outcome of the interaction between these simultaneous dynamic decision processes, and in particular the impact of the lags created by the perception formation process and the time to implement desired changes in capacity. These multiple delays may result in customers and service provider being out of step: customers walk away just as the service provider manages to bring extra capacity online.

One general conclusion from our simulations is that the facility is always better off if it updates its desired capacity faster than its current customers update their perception of waiting time (and thus by implication faster than the potential customers). More generally, the longer the process to increase capacity, the more important it is that companies be able to take fast decisions regarding desired capacity, as any delay in this part of the decision making process would prevent the company from taking advantage of potential expansion possibilities. In other words the action of the management team is an important factor in determining whether the company succeeds in achieving its growth potential.

Keywords: *System dynamics, behavioural queueing, simulation, capacity management*

Introduction and literature review

Traditionally queuing theory has focused on the design of service facilities, e.g. providing the appropriate capacity or the optimal number of service points. A good illustration of this is the broad literature on the optimal management of call centres (see for instance Gans et al (2003) for an overview of this work). There has been surprisingly little attention paid to the more behavioural aspects of the system, which are clearly important when dealing with people. Indeed, Gans et al. (2003) point out that the incorporation of human factors is a real challenge for the development of the next generation of queuing models for call centres.

The majority of papers in the queuing area consider the customer as an external factor, typically arriving following some well-known distribution, e.g. a Poisson process. The possibility that some of these customers might return to use the facility again if “treated well” or that dissatisfied customer might not return, and the potential impact of this on the arrival rate, has received little attention.

In this paper we develop a behavioural model in which customers come and go based on their perception of the waiting time (relative to other facilities) while managers gradually adjust the capacity of the facility based on their perception of demand. We investigate the outcome of the interaction between these simultaneous dynamic decision processes, and in particular the impact of the lags created by the perception formation process and the time to implement desired changes in capacity. These multiple delays may result in customers and service provider being out of step: customers walk away just as the service provider manages to bring extra capacity online. We begin by discussing the previous work leading up to our model.

There is a large literature in this area, going back to Naor’s (1969) seminal paper. While most of the early literature, generally referred to as queueing theory, is of a rather

technical nature, over the years a number of authors have started paying attention to customer behaviour, creating the field known as the psychology of queues. Early papers include Maister (1985) and Larson (1987). The inclusion of behavioural aspects in the study of service operations has attracted increasing attention, particularly in the marketing literature. Examples include Taylor (1994), Carmon et al (1995), Hui and Tse (1996), Kumar et al (1997), Zhou and Soman (2003) and Whiting and Donthu (2006). These papers focus on evaluating how waiting affects customer satisfaction and how the waiting process could be managed to minimise the resulting dissatisfaction. Few papers go one step further, asking the question: how will waiting time (dis)satisfaction affect customer loyalty (i.e. the likelihood of repeat business) and word of mouth (the impact on potential future customers).

Law et al (2004) and Bielen and Dumoulin (2007) have approached this issue from an empirical point of view. Law et al (2004) study the return frequency of a student population to various fast-food outlets. They conclude that while waiting time impacts customer satisfaction in all periods, it only has a significant impact on return frequency during the busy lunch period (p. 560). Bielen and Dumoulin (2007) find empirical support for their hypothesis that waiting time satisfaction moderates the effect of service satisfaction on loyalty. In particular, when waiting time satisfaction is high, loyalty will be high, whatever the overall service satisfaction level.

The objective of our work is to go a step further and improve our understanding of the long term consequences for a service provider of the impact of waiting time on customer retention and customer acquisition. We organize the paper as follows. After the overview of the queuing area provided above we discuss the context and provide examples to illustrate why it is essential to capture the behaviour of customers in more detail when investigating queuing systems. Next we develop a simple behavioural (feedback) model of this two stage queuing situation and analyse it using simulation. Finally, we generalize the key insights

derived from the analysis.

Context

When describing the model we use the term “queue” to denote the backlog of work. This could be a physical queue of customers, calls on hold, or a stack of files waiting to be dealt with. Similarly we use the term “waiting time” to denote the time from the moment the customer submits a request for service, until this service has been provided. In a manufacturing context this would be referred to as the sojourn time. Note that we will systematically phrase our discussion in terms of waiting time or response time, but the same arguments apply whenever service quality decreases as the utilisation rate of the resources increases. Examples include crowding, higher error rates due to stress, etc.

Consider a situation where customers choose a service provider and value a long term relationship. One example is the choice of a garage for care maintenance. Other examples include a subscription to a helpline for computer problems or choosing a tax accountant. In each of these situations, customers value being able to obtain “fast” access to the server, but the definition of “fast” depends on the type of service. For instance, one considers it acceptable to book an appointment for car maintenance several weeks ahead of time, or to have a two month delay between mailing income information to a tax accountant and receiving the filled-in tax form to sign. But for a computer helpline, the acceptable waiting time is of the order of minutes, if not seconds. For this reason, the theoretical model we develop is expressed in "time-units", without specifying whether these are seconds, hours or days.

Customers will tend to stay with their chosen service provider as long as they consider the waiting time acceptable compared to the generally accepted norm for this type of service. If they face increasing delays, they will consider changing provider. Different people will be

more or less patient, and the ease with which they will change provider also depends on their switching cost. For instance, selecting a new tax advisor implies investing time in identifying a suitable firm, negotiating terms, and explaining one's situation to the new advisor.

Even a service provider with a satisfactory service level will lose some customers as people move out of the area, stop driving, retire, etc. Similarly, as new people move into an area (or people acquire their first car or enter the job market), they will look for a service provider. Their choice will be influenced by the reputation of the different providers: the better the reputation of a provider, the higher the share of these new customers a provider can expect.

The service provider does have an incentive to invest in sufficient capacity to be able to provide a satisfactory service level, while avoiding overcapacity for obvious cost reasons. He thus needs to make two decisions: (i) determine his desired capacity and (ii) decide at what speed to adjust existing capacity if this differs from his desired capacity. For the latter decision, the speed of adjustment is likely to differ for capacity increases and decreases. Consider for instance a tax advisor. Increasing capacity means hiring new tax accountants, while decreasing capacity would imply lay-offs. The time required to implement the latter is likely to be influenced by legal constraints (e.g. a notice period) and the firm's policy regarding laying off people. The time required to increase capacity will be influenced among others by the availability of qualified people on the job market.

Concerning the desired capacity, the service provider is unlikely to know exactly how many customers he has. Taking again the tax advisor example, the firm may know how many people came the previous year, but there is no guarantee that they will come back this year. The most up to date information the firm has is the rate at which requests come in and the extent to which a backlog of work is building up. Combining this information with the desired response time will provide the firm with an estimate of how much capacity is required to

achieve an acceptable response time. We assume that the service provider is a small player, i.e. he is one among many competing firms. This implies that his behaviour, and in particular his response time, will have no noticeable impact on what the market considers to be acceptable.

Model Description

Below we present the model structure using feedback diagrams. The main advantage of feedback diagrams is that they provide an intuitive, non-technical understanding of the logic of the proposed model, and especially of the interaction of the different factors. This comes at the expense of some loss of mathematical accuracy, which is why we provide a full equation listing in Appendix A.

The model has two sectors representing the two main groups of actors: customers (composed of current and potential customers) and the management of the facility, whose behaviour determines the service capacity. We first describe the customer sector. The feedback loop in Figure 1 represents the behaviour of current customers.

In these diagrams, the arrow linking any two variables, x and y , indicates the existence of a causal relationship between x and y . The sign at the head of each arrow denotes the nature of the relationship as follows:

$$x \xrightarrow{+} y \Rightarrow \frac{\partial y}{\partial x} > 0 \quad \text{and} \quad x \xrightarrow{-} y \Rightarrow \frac{\partial y}{\partial x} < 0 \quad (1)$$

 Insert Figure 1 about here

Arriving customers join the queue and receive service in turn. Upon leaving the facility, the customer knows the waiting time he experienced and will use this experience to adjust his view of the average service time at this facility. The customer will not simply substitute this new value for his previous estimate, but instead update his estimate using this

new observed value. Such behaviour is often modelled using adaptive expectations where:

$$\begin{aligned} \text{Expected_Waiting_Time}_t \\ = \alpha \text{Experienced_Waiting_Time}_{t-1} + (1-\alpha)\text{Expected_Waiting_Time}_{t-1} \end{aligned} \quad (2)$$

where α is the weight the customer assigns to his new experience, and $(1-\alpha)$ the weight assigned to the previous expectation. Equivalently, one can interpret $\tau = 1/\alpha$ as the time required by current customers to update their expectation. The chosen value of α will vary with the frequency of use of the facility, the type of facility etc (Evans and Honkapohja, 2001). The comparison between this expected waiting time and what the customers consider to be an acceptable waiting time determines how satisfied the customers are with the queuing time at the facility:

$$\text{Actual_Customer_Satisfaction}_t = \frac{\text{Expected_Waiting_Time}_t}{\text{Acceptable_Waiting_Time}} \quad (3)$$

This satisfaction index then determines whether the customer will return at a later date (thus generating repeat business) or choose to leave the customer base. The rate at which customers leave the customer base is modelled as a nonlinear function of this level of satisfaction. Possible functional forms are discussed in more detail later on.

The outcome is a negative feedback loop: if the queue becomes too long for the customers' taste, more customers will leave, leading to a lower arrival rate and thus a shorter queue. However, the current model cannot increase the number of customers as there is no inflow of new customers when the satisfaction level is high.

The Feedback loop in Figure 2, which shows the behaviour of potential customers, captures the possibility of new customers joining the customer base. The logic is similar to that of the current customers, except that potential customers have no direct experience: they must rely on word of mouth to form an expectation of the waiting time at a specific facility. We thus assume a similar process of adaptive expectations to capture the gradual adjustment, but with an adjustment time longer or equal to that of the current customers.

Insert Figure 2 about here

These two loops capture the customer behaviour (customer loss and acquisition), and are similar in spirit to those of Agnew (1976), Haxholdt et al. (2003) and van Ackere et al. (2006), except for the fact that we allow for behavioural difference between current and potential customers, which is more realistic.

Both are negative feedback loops: when the perceived waiting times exceeds acceptable waiting time, more customers will leave and fewer will join, while the reverse holds when the perceived waiting time is below the acceptable waiting time. Consequently, over time the customer base, and thus the queue length, will tend to a value that will equalise expected and acceptable waiting times. Still, such a structure has the potential to result in fluctuations where periods with too many and too few customers alternate. The time required for convergence to a stable equilibrium (assuming such convergence does take place) will depend on the exact functional forms of the non-linear relationships and the time taken by current and potential customers to update their expectations.

Next we turn to the second part of the model which deals with capacity management of the service facility. Figure 3 illustrate this process.

Insert Figure 3 about here

For a given queue length, the service capacity determines the actual waiting time. Most service systems have at least two levels of capacity adjustment. One the one hand, there is usually potential to implement fast, limited adjustments to the existing capacity, for instance through the use of overtime, or opening additional tills in a supermarket. On the other hand, more fundamental adjustment of the capacity, e.g. new buildings, new IT systems, hiring and training of new employees etc., require considerable more time. It is this type of

capacity adjustments that we aim to capture in this part of the model.

We again use an adaptive expectations approach: we assume that the management of the facility assesses the queue length over some time period and compares this assessment to some notion of what a reasonable queue length is. This comparison yields some degree of satisfaction with the current situation, and depending on this management may decide that current capacity is satisfactory, too high or too low, depending on which they will decide to leave capacity as is, increase capacity or decrease capacity. The idea that management might consider the queue to be "too short" may seem surprising. Recall that the queue represents a backlog of work. If the backlog is such that management expects this to result in periods of underutilisation of capacity, they would wish to decrease capacity. So, although we do not explicitly model the economic aspects we capture the idea that a low backlog may be "uneconomical" and lead to a desire to reduce capacity.

The time required to increase or decrease capacity will depend greatly on the specific industry being modelled. For instance, it may be comparatively easy to increase the number of cashier desks in a store by modifying the physical layout, while obtaining planning permission for a new building could take years. Similarly, the time required to reduce capacity will depend on the context: for instance, in certain countries, a workforce reduction can be implemented within days, while in others this may require months of negotiations with labour unions.

The resulting change will bring the service capacity more in-line with the actual demand, yielding a third balancing loop (the "capacity" loop). Again, the presence of this additional balancing loop will not ensure that the system converges quickly to a stable equilibrium. This will be influenced by the time-lags involved: the time required to update the perception of required capacity, and the time required to increase or decrease capacity.

Finally, Figure 4 illustrates how the two parts of the model (customer behaviour and

capacity adjustment) interact. Our objective is to gain a better understanding of this interaction – the continuous adjustment of the customers to the waiting time and the simultaneous continuous capacity adjustments implemented by the management based on their perception of the situation. We are interested in understanding how the relative magnitude of the various lags in the model (both on the customer and the service provider sides) results in desirable or undesirable situations.

Insert Figure 4 about here

Experimental Strategy

Many classical linear queueing models allow for analytical solutions, however, when models become more realistic, and thus more complicated, the non-linearities result in the model being mathematically intractable, preventing the derivation of closed form solutions. Consequently, over the last decade, simulation and numerical analysis have increasingly become accepted as the most appropriate, if not the only way to analyse such models. The use of simulation has increased in many areas, including economics (Albin, 1998), organizational theory (Lomi and Larsen, 2001; Pritula et al., 1998) and medicine (Kollman et al., 2003). For a review of the use of simulation, see Zacharias et al. (2008) and Harrison et al. (2007).

To analyse the model, we use the well known approach of “carpet bombing” (Mosekilde and Larsen, 1988) i.e. we perform a systematic search across the multi-dimensional space determined by the model's key parameters. While the model was initially developed in a system dynamics framework, we implemented the model in C so as to be able to develop an automatic detection system for the long-term behaviour for each set of parameter values. While we acknowledge that in many cases the transition from the initial condition to a steady state can be an important element, we have at this stage chosen to focus on the long term outcome as indicator of the desirability of a given parameter combination.

We are interested in the types of long-term behaviour the model can generate with respect to the customer base.

These turn out to be “collapse”, i.e. rapidly going to a zero, converging to some constant value (which we for convenience split into two areas, low and high, depending on whether the convergence value is above or below the equilibrium value of the model) and exponential growth. We acknowledge that exponential growth is not a realistic long-term behaviour and a company with such a growth pattern would no longer fit our hypothesis that we are dealing with a small player in a large market. But faced with the trade-off between accepting this behaviour and introducing additional structure to prevent it (e.g. by modelling some form of saturation) we have chosen the former. Indeed, our objective is to shed light on the interaction between customer behaviour and capacity adjustments for a small player, not to study how this firm should change its behaviour in the light of sustained strong growth.

As stated above, our model contains five behavioural parameters, two relating to customer behaviour and three relating to capacity adjustment decisions. Keeping three of these constant and varying the remaining two, we use simulation and "carpet bombing" to create a variety of phase diagrams or basin of attraction maps which allow us to observe for which parameter combinations the different types of long-term behaviour mentioned above occur and thus identify desirable strategies for the management.

Initial values of state variables and parameter values

Table 1 summarises the parameter values (with their sensitivity range), the equilibrium values of the state variables and the initial values for the simulations. Our qualitative results do not depend on the choice of these values.

Insert Table 1 about here

Of more importance are the two nonlinear functions of the model, the *Impact on new customers* (which determines how the perception of potential customers affects the number of new customers) and the *Fraction joining Queue* (which determines the rate of repeat business) shown in Figure 5.

Insert Figure 5 about here

The *Impact on new customers* represents the impact of the perception of potential customers on how many new customers join. More specifically, the higher the ratio (denoted RP) between the perceived waiting time of potential customer and the acceptable waiting time (the market reference), the less attracted new customers are, and the fewer will join. When this ratio equals 1, (i.e. when the service provider is perceived to perform as the market expects), the impact is neutral (a value of 1). As this ratio increases, the impact becomes less than 1 (less new customers), and converges to 0 (i.e. no new customers) as the ratio goes to 2, i.e. when the perceived waiting time equals twice the market reference. When the ratio is below 1 (perceived waiting time is less than the reference), more customers will join. The maximum value of the function equals 2, implying that with excellent service the provider could attract at most twice his normal share of new customers.

The *Fraction joining Queue* represents the impact of the perception of current customers on how many of them choose to join the queue when requiring service (as opposed to looking for another provider and thus leaving the customer base). More specifically, the higher the ratio between the perceived waiting time of current customer and the acceptable waiting time (the market reference), the fewer will join the queue. When current customers' perception equals the market reference waiting time, 5% of existing customers are lost each period. This captures the idea that even if a firm meets market standards, there is a natural turnover rate due for instance to customers moving away. As service performance

deteriorates, an increasing fraction of current customers chooses to go elsewhere. For instance, for a value of 3 (i.e. the waiting time at the service provider is three times the market reference), only 25% of customers remain loyal, and for a value of 5, all choose to go elsewhere. Conversely, as this ratio decreases, the fraction of potential customers who join the queue converges to 1, i.e. all current customers needing service join the queue. While the exact shape of these functions affects the numerical results, the general observed patterns and qualitative conclusions are robust for any "reasonable" functional shape (i.e. inverse S-shaped functions with reasonable minimum and maximum values).

Simulation results

Base case

Before undertaking a more comprehensive analysis of the results we discuss the base case. To illustrate the dynamics we modify the initial conditions so as to take the model out of equilibrium. As starting point we assume that the service facility has successfully attracted some new customers, i.e. we set the initial value of the Customer base at 200 instead of the equilibrium value of 175 (see table 1). The choice of this initial value has no impact on the results as long as it is "reasonable". Figure 6 shows the behaviour of the most important variables of the model.

Insert Figure 6 about here

Starting with Figure 6a we observe that 200 customers is too much for the service capacity of 25 (Figure 6c), which results in an increase of the queue length (Figure 6b), and thus of the waiting time (Figure 6d). Management reacts by gradually increasing capacity (Figure 6c). The delays inherent in the reaction of customers and management create a damped oscillation: fluctuations gradually loose amplitude and we observe convergence to a

new equilibrium. The size and the period of the fluctuations depend on the model parameters. For other parameter values, the fluctuations can exhibit larger or smaller amplitude, and take more or less time to subside.

Furthermore, equilibrium is not always achieved. The model can exhibit four different styles of behaviour, as shown in Figure 7. We briefly characterize these to facilitate the general discussion of the results.

Insert Figure 7 about here

Figure 7a shows the case where the customer base, while oscillating, declines over time and the number of customers eventually goes to zero. This is the “death spiral” where the reactions of customers and management are out of phase, i.e. customer demand drops just when management starts to expand capacity and/or the demand increases as capacity is being closed down. In such cases, the unfortunate combination of lags makes the operation unsustainable: the customer base follows a downward trend, with smaller and smaller oscillations, until the last customer walks out of the door.

We can observe the opposite, a “virtuous circle” in Figure 7d, where the lags are in tune with each other, resulting in exponential expansion of service capacity, in parallel with increasing customer numbers. In this case we do observe increasing oscillations, (this also applies to the capacity and the queue length), which may not be desirable from an economic point of view. But, fundamentally, customers are “in phase”, which allows for continued expansion of the system. Still, as discussed before, in practice there would be external forces to halt, or at least slow down expansion at some point (e.g. limits of the potential market, investment constraints, etc).

The remaining two Figures (7b and c) show the third type of behaviour: a new equilibrium is reached. In such cases the interaction between the different lags is such that

after a while no further changes take place: customer expectations remain constant, and there is no desire to modify capacity. Recall that there are no structural differences between these four scenarios, only the (relative) speed of the adjustment of expectations and changes in capacity differ. Next we move on to a more detailed analysis of how the choice of the behavioural parameters affects model behaviour.

Simulation Experiments

To understand the interdependence and the interactions between the delays in the model and the resulting behaviour, we first attempt to create an understanding of the five dimensional space created by the three expectations formations (perceived waiting time for current customers (TPC) and potential customers (TPP) and the required time for management to estimate the queue (TPQ)) and the two capacity adjustment delays (TTI and TTD respectively). We present a number of illustrative simulations before discussing general insights. We assume that management's objective is to achieve sustained growth of the customer base, or at the very least avoid losing all customers. One could consider alternative objectives such as being able to guarantee a maximum wait. This would require capturing how waiting time varies of time, which is beyond the scope of this paper.

In Figure 8 we fix the time for potential customers to update their perception (TPP) at 4 and consider two values for the time current customers need to adjust their perception (TPC equals 1 and 4) and for the time it takes the manager to evaluate the queue and thus assess his desired capacity (TPQ equals 3 and 12). This yields four combinations, and for each of these we consider a full range of delays to increase or decrease capacity (respectively 0 to 12 and 0 to 18) using a carpet bombing technique.

Insert Figure 8 around here

When current customers update their expectations fast and management update their desired capacity rather fast ($TPC=1$ and $TPQ = 3$) the area with “attractive” outcomes (i.e. convergence to “above 50” and exponential growth) covers most of the phase space: only when the time to increase capacity is very long is there a risk of ending up in the “below 50” area. We do not discuss the relative desirability of the parameter combinations yielding attractive outcomes as this would require us to introduce cost and revenue aspects, which are beyond the scope of this paper. If current customers update their expectations more slowly ($TPC = 4$), the speed of capacity adjustment (increase and decrease) plays a key role in determining the outcome: the most undesirable solution of collapse (the customer base goes to zero) occupies almost a quarter of the phase space.

Next consider the case where management updates its expectations about required capacity very slowly ($TPQ = 12$). Whether current customers update fast or very fast ($TPC = 1$ or 4) has minimal impact as in both cases collapse takes place over most of the space (respectively 70% and 85% of the space). This illustrates the interdependence of the different reaction delays and thus the need to consider them jointly: if management needs a long time to determine desired capacity, there are only very few parameter combinations that yield a desirable outcome, making the system challenging to manage. Conversely, if management updates its expectations faster, there is more room to manoeuvre, especially if current customers react not too fast ($TPC = 4$): by adjusting the time to increase and decrease capacity management can impact the outcome significantly.

In Figure 9 we choose a medium speed for the updating of management expectations ($TPQ = 6$) and consider two values for the time to increase and decrease capacity (TTD equals 1 and 9 and TTI equals 3 and 6 respectively). For each combination we let the updating times of both current and potential customers vary from 0 to 24. It is logical to assume that current customers have more up to date information than potential customers, and thus update their

expectations on average faster. For this reason we only consider parameter combinations above the 45° line in the phase diagrams.

Insert Figure 9 around here

One observes at once that the combination of slow capacity increase and very slow capacity decreases ($TTI = 6$ and $TTD = 9$) is highly undesirable as convergence to zero takes place almost everywhere, the only exception being a situation where current customers update their expectations very fast, and potential customers update theirs either very fast or very slowly. This situation occurs repeatedly and can be interpreted intuitively as follows. If potential customers update very slowly they are de facto not reacting to short term fluctuations in waiting time and will thus not be the cause of a sudden unexpected inflow of arrivals which would push waiting times up and drive away current customers. Similarly, if they update their expectations very fast, they will react quickly to changes in waiting time, in phase with the current customers, and thus increase the arrival rate when waiting times are short and there thus is plenty of capacity to accommodate them without driving existing customers away. It is for intermediate values that they react "out of sink", which creates problems for the system. At the other extreme ($TTI = 3$, $TTD = 1$), the situation is much more favourable: whatever the customer behaviour, the system converges to "above 50".

In case of a rather long time to decrease capacity ($TTD = 9$) combined to a short time to increase capacity ($TTI = 3$) the almost vertical borders between the different regions imply that, given the reaction times of current customers, the reaction time of potential customers has little or no impact. This is interesting as management clearly has more scope to influence the reaction time of current customers (e.g. through better information) than that of potential customers; we will discuss this further on in more detail. In this situation it is desirable for the facility to have customers who react either very fast, or slowly: intermediate values imply

convergence to 0.

Finally, when capacity decreases are easily implemented ($TTD = 1$) and capacity increases rather slowly ($TTI = 6$) we again observe the limited impact of the reaction time of the potential customers. If current customers update fast, the reaction time of the potential customers has no impact and the system converges to "above 50". If current customers take long to update their expectations, the situation is less desirable: unless potential customers are very slow (in which case the system converges to "below 50"), the company will collapse.

Discussion of Simulation Results

In the previous section we discussed in detail a few selected cases to illustrate the type of results yielded by the model. Next we summarise the key insights obtained from the full set of simulations as it is unrealistic to discuss in detail every single parameter combination. We start our discussion with recommendations for the manager depending on the characteristics (speed of expectation updating) of his current and potential customers. We distinguish two cases, depending on whether or not there is a significant difference in updating speed between current and potential customers.

Current and potential customers update their expectations at similar speed

Table 2 summarises the main results relating to desirable management response. Whatever the reaction time of customers, if management updates its expectation of required capacity fast, they need to be able to increase capacity fast, and the speed at which capacity can be decreased plays a minor role. But if management elects to update capacity requirements more slowly (e.g. because of significant investment requirements or high sunk costs), they must be able to increase capacity very fast, and decrease capacity also quite fast to avoid going out of business. The detailed simulation results show that in all cases,

management is better off if they update expectations fast and are able to increase capacity fast. This raises a more tricky issue though: updating capacity fast implies a higher risk of error, and thus inappropriate capacity adjustment decisions which might need to be reversed at high cost.

Overall, the management strategy is relatively predictable in this case as could be expected when both customer segments respond in the same way. The main choice for the management is between the time they take to assess required capacity and the time it takes to implement an increase in capacity. If management assesses their capacity needs more slowly, the capacity increase needs to be very fast to avoid losing significant amounts of customers or growth potential. The time required to increase capacity is situation specific as a supermarket can relatively quickly (at least up to a certain point) open more checkouts, particularly if there are employees around who can be redeployed, but even physically adding more checkouts or training employees is relatively fast compared to say the time required to build a new factory.

Although our focus is on the gain or loss of customers (as opposed to the potential cost of unused capacity), one should not overlook the impact of the speed of capacity decrease due to its interaction with management's other decision parameters. Indeed, if the reaction time to decrease capacity is very short, any slack will be eliminated quickly, increasing the risk of a capacity shortage if more customers show up. Consequently, from a market share point of view, management will be better off closing down excess capacity more slowly, unless they are able to ramp up capacity at sufficient speed, in which case the speed of capacity decrease plays no significant role in the acquisition and/or retention of customers.

Insert Table 2 around here

Current customers update expectations significantly faster than potential customers

Here we consider the situation where current customers are significantly better informed than potential customers. This would be the case if current customers use the service regularly, or if there are barriers to communication with potential customers. For instance, in certain countries legal and medical services are subject to severe restrictions on the right to advertise, so information spreads by word of mouth, implying that potential customers will be less well informed than current customers.

The desirable management responses are summarized in Table 3. If management update their expectations fast, the speed of capacity adjustment (increase or decrease) has no impact, with the exception of the first case: given the slow reaction time of potential customers, downward adjustment of capacity should also take place more slowly. More generally, if management decides quickly on desired capacity, they can afford to take some time to implement the required adjustments. But recall the drawback of fast decisions regarding required capacity: there is a risk of incorrect decisions that might lead to a volatile (and expensive) capacity adjustment strategy.

The situation is more complicated when management updates expectations slowly. Such a strategy requires them to be able to increase capacity fast; otherwise there is a high risk of collapse. The closing down decision is trickier, in particular when current customers react fast. In this case management must either increase capacity fast, i.e. in line with the speed of reaction of the current customers, or very slowly (i.e. basically not reacting to the behaviour of current customers). Intermediate values would result in the capacity adjustments being "countercyclical" to the behaviour of the current customers. In other words, disregarding the cost aspect, it is optimal to always have sufficient capacity ready at the moment it is needed. To achieve this, one needs to react either very quickly or very slowly. The former implies having always the right amount of capacity, and the latter corresponds to always having plenty of capacity available. Anything in between these two extremes is less

desirable.

Insert Table 3 around here

Finally, we consider the interrelationship between expectation updating of potential and current customers, and managements updating of expectations as shown in Table 4. Customers' speed of reaction is based on the time it takes them to update their expectations, which in turn depends on their available information. As mentioned before, current customers by definition have more timely information, and thus can update their expectations faster. Additionally, managers are much more likely to be able to influence (increase) the amount of information available to their current customers than the information available to potential customers. Consequently, they have some ability to reduce their existing customers' reaction delay. Table 4 analyses if and when this is a desirable initiative for managers.

If potential customers update their expectations fast, we are in a situation where ample information is available, so current customers will also update fast and management cannot impact this aspect. In this situation they are better of updating their own expectations regarding desired capacity quickly, which is why we indicate "No Choice" in the "Preference" column. In the intermediate case, management could influence the amount of information available to speed up current customers' updating process. If they do, it is desirable for them to update their assessment of capacity needs at a medium speed, while if they don't, they should update their assessment fast. Given a choice, they are better of speeding up the customers' reaction time.

When potential customers have limited information and thus a slow updating process, the situation is less obvious. The preferred updating speed for current customers and the corresponding rate at which managers should update their assessment of desired capacity is influenced by whether or not management is able to increase capacity quickly. If they are

unable to do so, the safe choice is to maximise information available to current customers to shorten their reaction time, and update the assessment of capacity needs fast. If management is flexible when it comes to increasing capacity, they are better off limiting available information, thus slowing down the current's customers' reaction, while updating their assessment of capacity fast.

Insert Table 4 around here

General insight from the simulations

One general conclusion from our simulations is that the facility is always better off if it updates its desired capacity faster than its current customers update their perception of waiting time (and thus by implication faster than the potential customers). Considering for instance a supermarket, this implies that the management should increase the number of cashiers at the first signs of crowding, rather than wait until there are significant queues and customers become dissatisfied. Similarly, on a different time-scale, there is an incentive for internet providers to be pro-active in increasing bandwidth before demand exceeds supply, at the risk of having over-capacity if the expected increase does not fully materialise.

Next, let us focus on the special case where the main capacity resource is labour, and consider situations where reducing capacity is a lengthy process. This would be the case for instance in countries with strong labour protection laws. To stay in the desirable part of the state space under such conditions, companies must either update their expectations about required capacity rather quickly, or be able to increase capacity quite fast. The former leaves the company highly exposed to misinterpreting market trends (i.e. reacting to a short-term fluctuation as if it were a long term trend), and thus adjusting capacity in the wrong direction. The latter may be difficult, if not impossible, if one is facing a tight labour market, or if a lengthy company-specific training is required. There is also a psychological barrier:

management aware of the difficulty to lay off people are bound to be cautious when it comes to hiring. This may lead to a vicious spiral where the company does not dare add capacity at the right time, acquires a reputation for long waits, and ends up out of business.

More generally, the longer the process to increase capacity (be it due to a tight labour market or the delays inherent in implementing investment decisions), the more important it is that companies be able to take fast decisions regarding desired capacity, as any delay in this part of the decision making process would prevent the company from taking advantage of potential expansion possibilities. In other words the action of the management team is an important factor in determining whether the company succeeds in achieving its growth potential.

Conclusion and Discussion

By introducing feedback into the traditional model of queuing, we no longer consider it as a “linear” process: we allow satisfied customers to return to use the facility and disappointed customers to stay away. Furthermore, we explicitly account for the difference in access to information between existing and potential customers, which implies that these two groups have different time constants to update their expectations about the waiting time at the facility. Finally, we have introduced two managerial elements into the model: management can decide how fast they update their assessment of required capacity and there are time lags to actually implement capacity increases and decreases once the need for these has been established.

Our analysis of the interrelationship between these five parameters has led to a number of general conclusions, some of which are expected while others are less intuitive. A manager who reduces capacity fast when demand seems to be decreasing can find himself in major trouble when things pick up again, especially if his time to increase capacity is relative long.

In particular, one might end up in a situation where capacity adjustments are out of phase with changes in customer demand: the desired capacity reduction is being implemented as demand starts to increase and vice versa. Situations like these have been observed repeatedly, over many decades, in both the chemical and computer memory chips industries: demand and capacity are often out of phase, causing periods of high prices to alternate with periods of very depressed prices. These industries are characterised by expensive, and thus slow, changes in capacity, in both directions.

Our simulations indicate that the relative magnitude of time to increase or decrease capacity has an important impact on the evolution of the customer base. This is an important observation as in many cases management's ability to influence these delays (and particularly their ability to shorten them) is very limited. Our main observations can be summarised as follows.

First, increasing capacity slowly can have undesirable consequences. These include:

(i) Being comparatively faster at increasing capacity than at decreasing capacity will generally yields satisfactory results in the short run but can backfire in the long run, as was illustrated in Figure 7a: the initial expansion soon reverses, and in the longer term the customer base collapses.

(ii) A company with the right "balance" between opening and closing capacity will be able to develop a stable customer base, but the definition of the "right balance" depends on the speed of reaction of the current and potential customers, as well as on the rate at which the company evaluates capacity needs, as was illustrated in the discussion of tables 2 and 3.

(iii) A strategy based on responding quickly to a perceived need to increase capacity, combined with a conservative attitude towards decreasing capacity looks attractive, but could be quite dangerous, especially for a small firm. Indeed, at some point a fast growing company may find itself unable to continue to grow resources sufficiently fast, leading to an increase in

the time to increase capacity. This would shift the company quickly into the collapse region, from where an escape is close to impossible. This trap is particularly dangerous for companies who need scarce resources (e.g. highly trained experts) to grow.

The third managerial parameter, the time to update expectations about required capacity, adds another dimension to the problem, especially if management has limited ability to modify the time it takes to adjust capacity. If management takes too long to update expectations about required capacity, a situation where capacity is out of phase with the changes in customer demand is likely to occur. At best this will result in a situation with significantly less customers than could be achieved, and at worse it will lead to a total collapse as customers fail to return.

We have also discussed the importance of the speed at which current and potential customers update their expectations and the fact that managers should adapt their decision making process to customer behaviour. This is particularly true as managers' ability to influence customer behaviour (e.g. by making information more readily available), is limited, especially with respect to the potential customers.

The model we have analysed has many simplifications and thus limitation. Our aim has been to go beyond the classical queueing model by introducing feedback, both in terms of customer behaviour and capacity planning. This has enabled us to analyse the interplay between customer behaviour (expectation formation resulting in joining or leaving decisions) and managerial decision making (capacity planning and implementing adjustments). Despite the many simplifications, the presence of feedback and nonlinearities prevents the derivation of any kind of closed form solution, thus making it necessary to resort to simulation to analyse the model.

One limitation is that, while we assume that the service provider is a small player in a large market, we have chosen not to incorporate limits to growth into the model, other than

those deriving from the managerial decisions regarding capacity planning and adjustment, thus opening the door to exponential growth for certain parameter combinations. Including such constraints would make the model even more complex, making it harder to draw any insights, without adding to its usefulness.

One such constraint would be to model the fact that exponential growth of a firm implies that this firm becomes a major player. Consequently a change in the firm's response time would affect the market reference (the waiting time customers consider acceptable), which would thus no longer be an exogenous parameter. This would add another important feedback to the model, and including it in a meaningful way would require us to address another limitation of the current model: the static nature of the different delays. Indeed, we allow neither the customers nor the manager to change behaviour over time. But, if a firm is allowed to evolve from being a small player to become a market leader, one must also allow for a change in behaviour (different approach to capacity management) by this firm and a change in the reaction time of the potential customers, as information is more readily available about the performance of a market leader than about the performance of a small firm. Including this latter aspect would require us to move the analysis to a different level, i.e. we might need to model customers at the individual level. Such an extension is beyond the scope of our current work.

At this stage we focus exclusively on behaviour, so the model does not include economic considerations. Adding this aspect would be the first obvious direction in which to extend the model, i.e. move to a multi-objective decision problem where the firm trades off growth (in terms of the size of the customer base) and profitability. This would also enable us to look at trade-off between short-term profit and sustainable long term growth.

Acknowledge

The work leading to this paper was made possible by the support received from the Swiss

National Science Foundation, Research Grant Number 100012-116564 / 1.

References

- Albin, P.S. 1998. Barriers and bounds to rationality. Princeton, NJ: Princeton University Press
- Bielen, F., N. Dumoulin. 2007. Waiting time influence on the satisfaction-loyalty relationship in services. *Managing service quality*, 17(2), 174-193.
- Carmon, Z., J. G. Shanthikumar, T. F. Carmon. 1995. A psychological perspective on service segmentation models: The significance of accounting for consumers' perceptions of waiting and service. *Management Science*, 41(11), 1806-1815.
- Evans G.W., Honkapohja, S. 2001. Learning and expectations in macroeconomics. Princeton University Press: NJ.
- Gans, N, G. Koole and A. Mandelbaum. 2003. Telephone Call Centers: Tutorial, Review, and Research Prospects, *Manufacturing & Service Operations Management* 5:79–141
- Harrison, R.J., Z. Lin, G.R. Carroll, and K.M. Carley. 2007. Simulation Modeling in Organizational and Management Research. *Academy of Management Review*, 32: 1199-1228.
- Haxholdt, C., E. R. Larsen, A. van Ackere. 2003. Mode Locking and Chaos in a Deterministic Queueing Model with Feedback. *Management Science*, 49(6), 816-830.
- Hui, M. K., D. K. Tse. 1996. What to tell consumers in waits of different lengths: An integrative model of service evaluation. *Journal of Marketing*, 60(2), 81-90.
- Kollman, K., Miller, J.H., Page, S.E. 2003. Computational models in political economy. Cambridge, MA: MIT-Press
- Kumar, P., M. U. Kalwani, M. Dada. 1997. The Impact of waiting time guarantees on customers' waiting experiences. *Marketing Science*, 16(4), 295-314.
- Larson, R., 1987. Perspectives on queues: social justice and the psychology of queuing.

- Operations research, 35(6), 895-905.
- Law, A. K. Y., Y. V. Hui, X. Zhao. 2004. Modeling repurchase frequency and customer satisfaction for fast food outlets. *International journal of quality & reliability management*, 21(5), 545-563.
- Lomi, A., Larsen, E.R. 2001. *Dynamics of organizations*, Cambridge, MA: AAAI Press / MIT-Press
- Maister D., 1985. "The psychology of waiting lines" in *The service encounter*, J. Czepiel, M. Solomon, C. Surprenant (Eds.), Lexington, MA: Lexington Books.
- Mosekilde, E., Larsen, E.R. 1988 Deterministic chaos in the beer production-distribution model, *System Dynamics Review*, 4, 131-147
- Naor, P. 1969. On the regulation of queue size by levying tolls. *Econometrica*, 36(1), 15-24.
- Prietula, M.J., Carley, K.M., Grasser, L. 1998. *Simulating organizations* Cambridge, MA: AAAI Press/MIT-Press
- van Ackere, A., C. Haxholdt, E. R. Larsen. 2006. Long and short term customer reaction: a two-stage queueing approach. *System Dynamics Review*, 22(4), 349-369.
- Whiting A., N. Donthu. 2006. Managing Voice-to-Voice Encounters: Reducing the Agony of Being Put on Hold. *Journal of Service Research*, 8(3), 234-244
- Taylor, S. 1994. Waiting for Service: The relationship between delays and evaluations of service. *Journal of Marketing*, 58(2), 56-69.
- Zacharias, G.L., MacMillan, J., Van Hemel, S.B. 2008 *Behavioural modeling and simulation*. The National Academies Press.
- Zhou, R., D. Soman. 2003. Looking back: Exploring the psychology of queueing and the effect of the number of people behind. *Journal of Consumer Research*, 29(4), 517-530

FIGURE 1

Retention of Current Customers

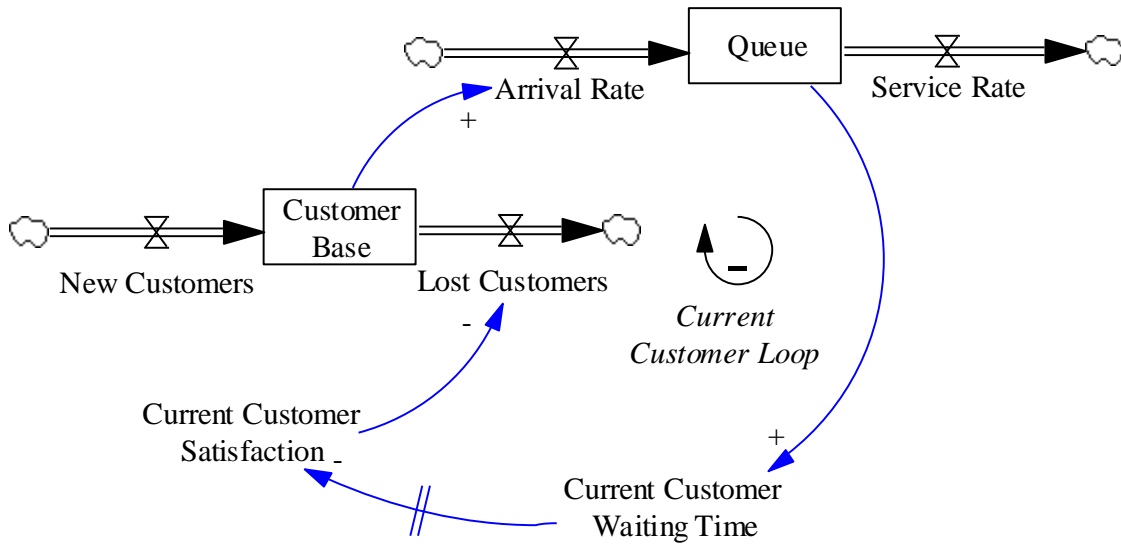


FIGURE 2

Behaviour of Potential Customers

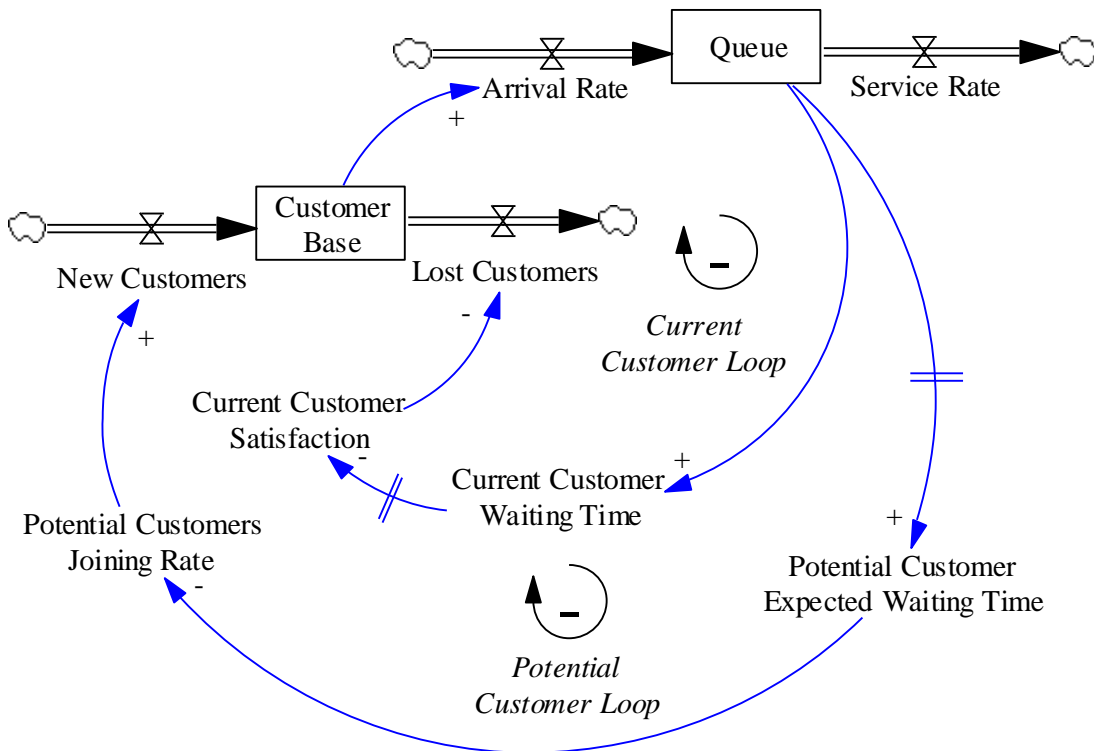


FIGURE 3
Managerial Behaviour: Capacity Planning and Adjustment

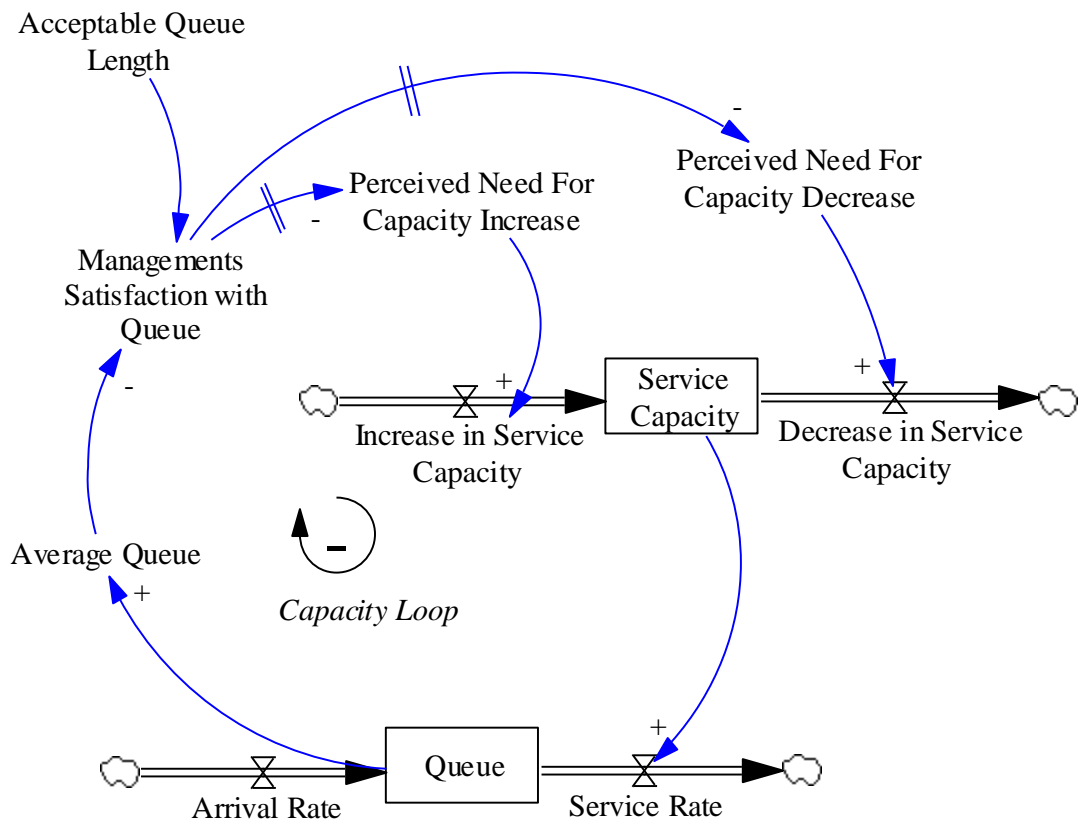


FIGURE 5

Impact of perceived waiting time on the customer base

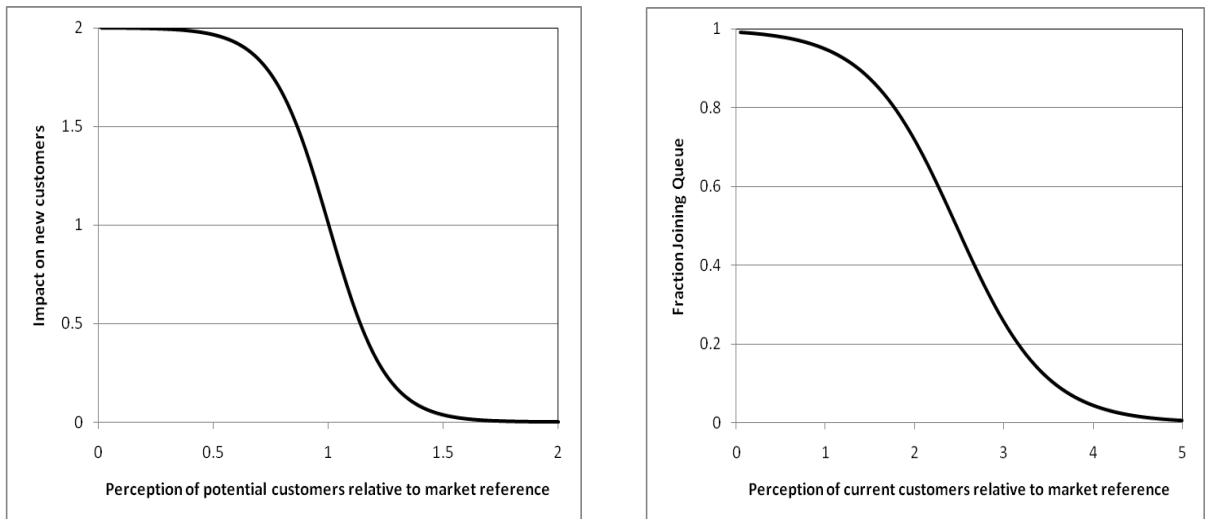


FIGURE 6

Illustration of Transient Behaviour in the Base Case

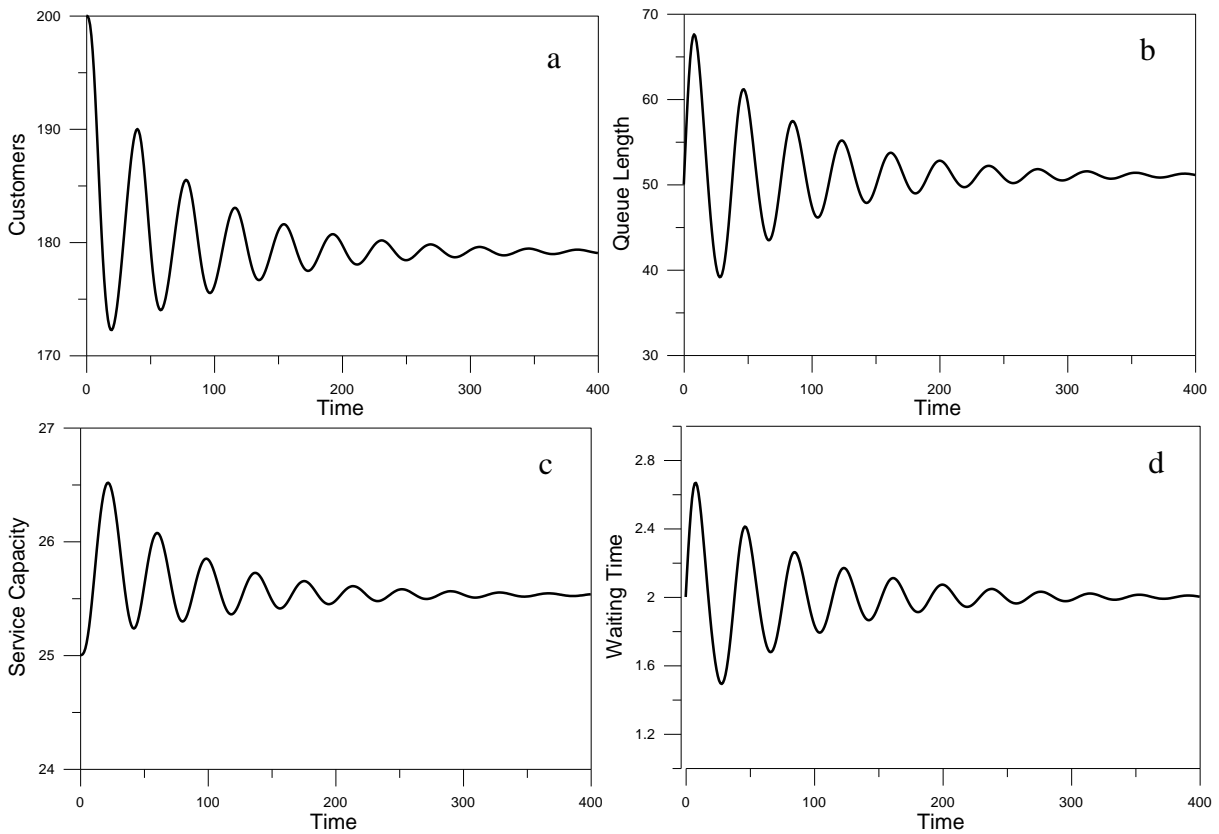


FIGURE 7

Illustration of the four types of model behaviour: Evolution of the customer base

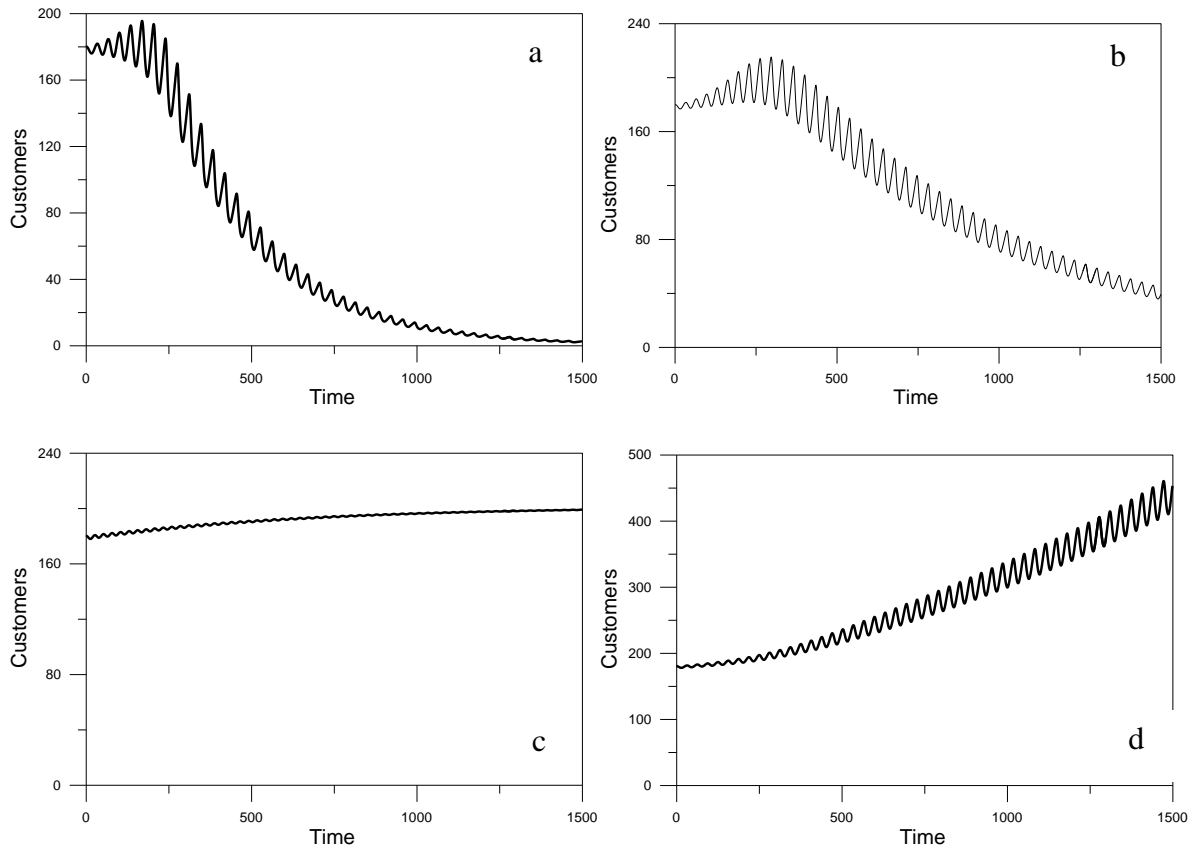


FIGURE 8

Illustrative simulation results for the case with relatively fast Potential customers ($TPP = 4$):

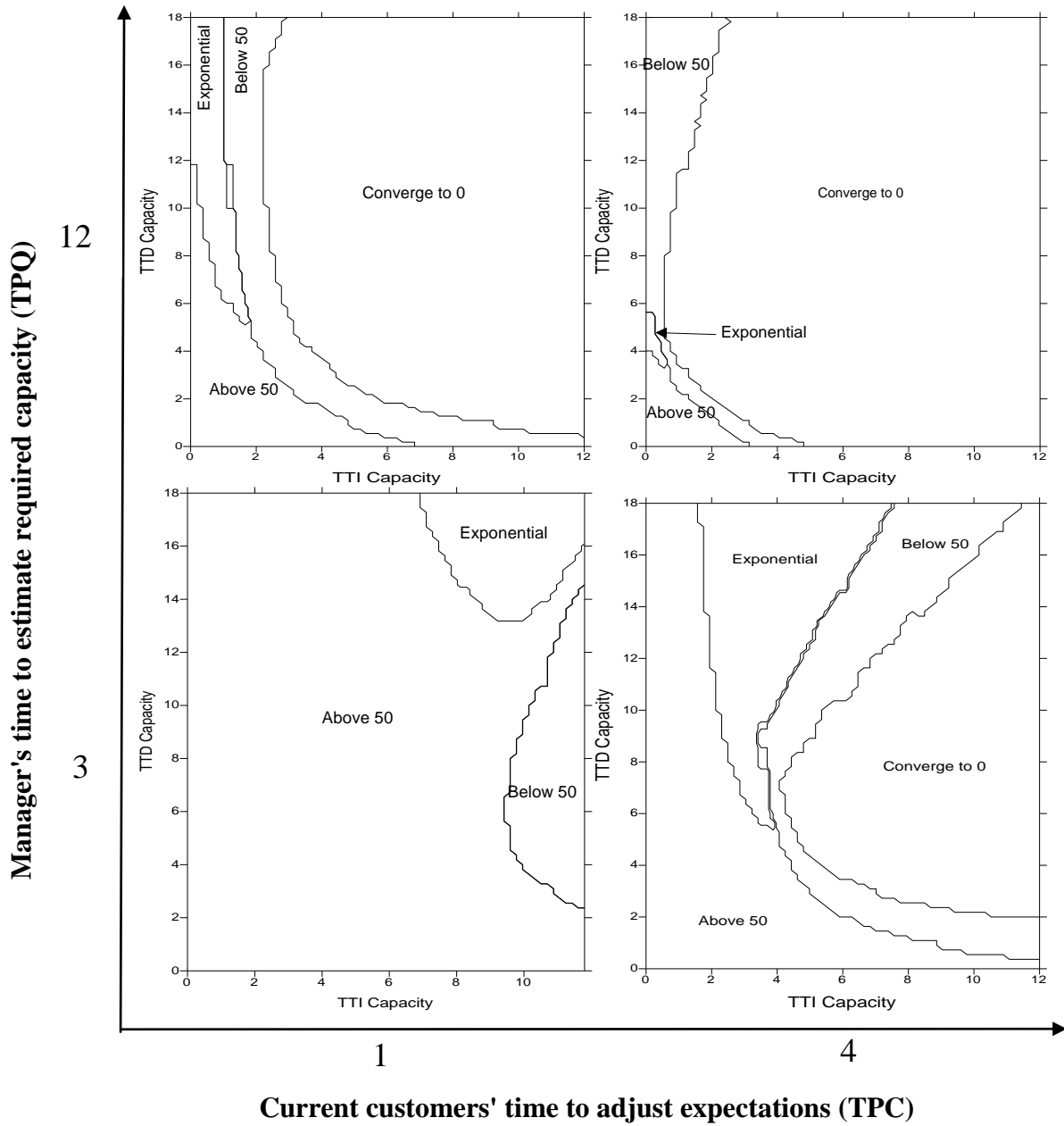


FIGURE 9

Illustrative simulation results: intermediate capacity assessment time (TPQ = 6)

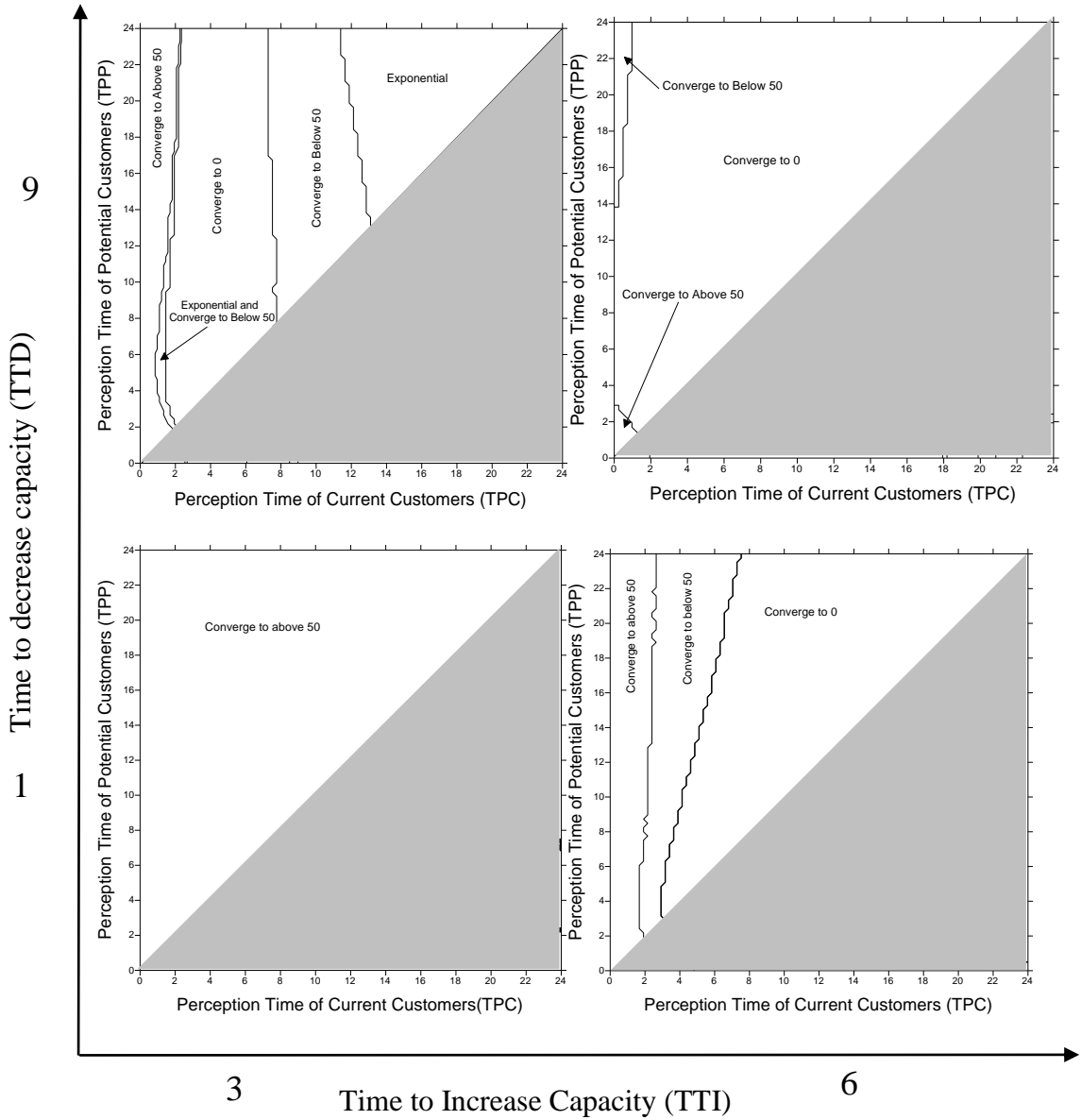


TABLE 1.
Initial values of state variables and parameter values

State variables	Equilibrium values	Unit
Customers	175	People
Queue	50	People
Average Queue	50	People
Service Capacity	25	People / Time unit
Perceived Waiting time current customers	2	Time unit
Perceived Waiting time potential customers	2	Time unit

Exogenous parameters	Value	Unit
Visits per time unit	0.15	1/Time unit
Minimum service Time	0.25	Time unit
Market reference waiting time	2	Time unit
Initial Customers	180	People
Normal joining rate	0.0075	1/Time unit

Time constants	Sensitivity range	Unit
Time to perceive queue length (TPQ)	1 - 12	Time unit
Time to increase capacity (TTI)	1 – 12	Time unit
Time to decrease capacity (TTD)	1 – 18	Time unit
Perception time of current customers (TPC)	1 - 24	Time unit
Perception time of potential customers (TPP)	1 - 24	Time unit

TABLE 2
Desirable management response when current and potential customers update their expectations at similar speed

Existing Customers Expectation Update	Potential Customers Expectation Update	Desirable Management Responses		
		<i>Management Expectation Update</i>	<i>Desired Time to Increase Capacity</i>	<i>Desired Time to Decrease Capacity</i>
Slow	Slow	Fast	Short	No Impact
		Slow	Very Short	Short
Medium	Medium	Fast	Short	No Impact
		Slow	Very Short	Short
Fast	Fast	Fast	Short	No Impact
		Slow	Very Short	Short

TABLE 3.

Desirable management response when current customers update their expectations faster than potential customers

Existing Customers Expectation Update	Potential Customers Expectation Update	Desirable Management Responses		
		<i>Management Expectation Update</i>	<i>Desired Time to Increase Capacity</i>	<i>Desired Time to Decrease Capacity</i>
Medium	Slow	Fast	No Impact	Long
		Slow	Short	No Impact
Fast	Slow	Fast	No Impact	No Impact
		Slow	Short	No Intermediate Values
Fast	Medium	Fast	No Impact	No Impact
		Slow	Short	No Intermediate Values

TABLE 4

Impact of the speed of perception of current and potential customers on the speed at which managements should update its perception of desired capacity

Potential Customers Expectation Updating	Existing Customers Expectation Updating	Management Expectation Update	Preference
Fast	Fast	Fast	No Choice
Medium	Fast	Medium	X
	Medium	Fast	
Slow	Fast	No Impact if TTI short else Fast	Safe Choice
	Medium	Fast or Medium	
	Slow	Fast	If TTI short

Appendix A. Equation listing

$$\text{Outside Dissatisfaction} = \frac{\text{Outside Perceived Waiting Time}}{\text{Market Reference Waiting Time}}$$

$$\text{Inside Dissatisfaction} = \frac{\text{Inside Perceived Waiting Time}}{\text{Market Reference Waiting Time}}$$

$$\text{Impact on new Customers} = f[\text{Outside Dissatisfaction}]$$

$$\text{New Customers} = \text{Normal Joining Rate} \times \text{Customers} \times \text{Impact on new Customers}$$

$$\text{Fraction Joining Queue} = g[\text{Inside Dissatisfaction}]$$

$$\text{Potential Arrivals} = \text{Customers} \times \text{Visits per Time Unit}$$

$$\text{Lost Customers} = \text{Potential Arrivals} \times (1 - \text{Fraction Joining Queue})$$

$$\text{Arrival Rate} = \text{Potential Arrivals} \times \text{Fraction Joining Queue}$$

$$\text{If } \frac{\text{Queue}}{\text{Minimum Service Time}} \geq \text{Service Capacity: } \text{Service Rate} = \text{Service Capacity}$$

$$\text{else: } \text{Service Rate} = \frac{\text{Queue}}{\text{Minimum Service Time}}$$

$$\text{Required Capacity} = \frac{\text{Average Queue}}{\text{Market Reference Waiting Time}}$$

$$\text{If } (\text{Required Capacity} - \text{Service Capacity}) > 0: \text{Capacity Increase} = \frac{\text{Required Capacity} - \text{Service Capacity}}{\text{TTI Capacity}}$$

$$\text{else: } \text{Capacity Increase} = 0$$

$$\text{If } (\text{Required Capacity} - \text{Service Capacity}) < 0: \text{Capacity Decrease} = \frac{\text{Required Capacity} - \text{Service Capacity}}{\text{TTD Capacity}}$$

$$\text{else: } \text{Capacity Decrease} = 0$$

$$\text{Current Waiting Time} = \frac{\text{Queue}}{\text{Service Rate}}$$

$$\frac{d\text{Customers}}{dt} = \text{New Customers} - \text{Lost Customer}$$

$$\frac{d\text{Queue}}{dt} = \text{Arrival Rate} - \text{Service Rate}$$

$$\frac{d\text{Service Capacity}}{dt} = \text{Capacity Increase} - \text{Capacity Decrease}$$

$$\frac{d\text{Average Queue}}{dt} = \frac{\text{Queue} - \text{Average Queue}}{\text{TTE Average Queue}}$$

$$\frac{d\text{Inside Perceived Waitng Time}}{dt} = \frac{\text{Current Waiting Time} - \text{Inside Perceived Waitng Time}}{\text{Time to Perceive In}}$$

$$\frac{d\text{Outside Perceived Waitng Time}}{dt} = \frac{\text{Current Waiting Time} - \text{Outside Perceived Waitng Time}}{\text{Time to Perceive Out}}$$