Att använda kömodeller för att mäta kapacitet i ett system

Björn Lantz[1] *och Peter Rosén*[2]

1. *Chalmers tekniska högskola*
   412 96 GÖTEBORG
   031- 772 13 81
   bjorn.lantz@chalmers.se

2. *Handelshögskolan vid Göteborgs universitet*
   Box 610
   405 30 Göteborg
   031-786 44 87
   peter.rosen@handels.gu.se

SAMMANFATTNING

Den sanna kapaciteten i ett produktions- eller servicesystem är ofta svår att skatta med tillräcklig precision. De två huvudsakliga metoderna för kapacitetsmätning som beskrivs i litteraturen - mätning baserade på tekniska/fysiska parametrar och mätning baserad på tidsstudier - karakteriseras båda av såväl konceptuella som praktiska nackdelar. I detta papper analyseras hur kapaciteten i ett system istället kan mätas med hjälp av kömodeller, vilket eliminerar vissa större svagheter med de två andra metoderna. Den grundläggande idén är att använda det systemteoretiska samband som gäller mellan ankomsttakt, kapacitet och kötid: eftersom den takt med vilken jobb/kunder ankommer och den tid de befinner sig i kö ofta kan observeras objektivt kan dessa parametrar under vissa omständigheter användas för att indirekt mäta kapaciteten i det aktuella systemet. De primära fördelarna jämfört med de två traditionella metoderna för kapacitetsmätning är att den kömodellbaserade metoden är helt okänslig för förhållandet mellan designad och effektiv kapacitet och att den inte genererar några beteendemässiga problem eftersom den inte kräver någon direkt observation av systemet. Validitetsaspekter diskuteras också i papperet.

1. INTRODUCTION

Capacity planning is one of the major fields of operations management. Ensuring that operations can meet current and future demand cost-effectively is a fundamental task of operations managers (Horngren et al. 2012; Slack et al. 2010). This applies to both manufacturing firms and service organizations. For service organizations, the task is complicated by the fact that services are typically intangible, perishable, and cannot be stored. Services are often produced (capacity) and consumed (demand), simultaneously, at the same place and time and when in contact with consumers. Storage is not an available option to balance capacity and demand or to improve capacity utilization and service levels (Klassen and Rohlender 2001; Ng et al. 1999; Prajogo 2006; Sasser 1976). Therefore, capacity planning for service organizations is considered a more complex and complicated task than capacity planning for manufacturing firms (Adenso-Días et al. 2002; Brown et al. 2005;

Corsten and Stuhlmann 1998; Klassen and Rohlender 2001; Ng et al. 1999; Shemwell and Cronin 1994). However, for both types of organizations, developing a reliable methodology for measuring capacity is important because accurate data input is a prerequisite to effectively fulfilling capacity planning (Bamford and Chatziaslan 2009; Elmaghraby 1991: Fitzsimmons and Fitzsimmons 2008; Larsson 2013; Pullman and Rodgers 2010; Sasser 1976).

Queueing models are often used by applied researchers to characterize and analyse different types of systems in different types of industries, for example, call centres (Gans et al. 2003; Koole and Mandelbaum 2002), healthcare (Green 2006; Lakshmi and Iyer 2013; Palvannan and Teow 2012;), hospitality and tourism (Hwang et al. 2010; Pullman and Rodgers 2010), manufacturing (Govill and Fu 1999; Papadopoulos and Heavey 1996: Rao et al. 1998), retailing (Defraeye and Nieuwenhuyse 2016; Kesavan and Mani 2015), seaports (Canonaco et al. 2007; Dragovic et al. 2006; Shabayek and Yeung 2000; Stahlbock and Voss 2008), and telecommunications (Bruneel et al. 2014). In these industries, queueing models are applied to capacity planning and control. Queueing analysis is, for example, used to determine the capacity required by a service process to achieve an acceptable service level (i.e. waiting time) and to evaluate how changes in demand or the service process affect the queue lengths and waiting times (Brown et al. 2005; Green et al. 2006; Karvonen et al. 2004; Kim et al. 1999; Palvannan and Teow 2012).

Most comprehensive textbooks on the subject of operations management include the fundamentals of queueing theory (Balakrisnan et al. 2014; Fitzsimmons and, Fitzsimmons 2008; Krajewski et al. 2016; Lantz 2015; Slack et al. 2010). The typical textbook approach introduces the problem with queues that arise in many types of service industries as well as in manufacturing when the arrival of customers (or jobs) is a stochastic process and the service (or production) time is a random variable. The textbook approach is a straightforward presentation of the formulas used to calculate different operating characteristics for several different types of basic queueing systems. Although many textbook authors do mention the conceptual logic on which the formulas rely, the explicit derivation of the formulas is typically suppressed (as it should be because the math required is of a relatively advanced level); however, useful relationships to the context, such as Little’s law, are often mentioned (Little 1961, 2011; Taha 1981).

The formulas used to calculate the operating characteristics for a system assume that the arrival process and the service time are stochastic and characterized by specific statistical distributions with known parameters. In other words, given specific assumptions regarding the capacity of the system and the customer arrival process, we can use the formulas to compute, for example, the expected queueing time for a customer or the expected number of customers in the queue. However, the arrival process and the operating characteristics that the queueing formulas are used to compute, in most cases, can both be measured objectively in an empirical sense (we can observe the actual queue itself as well as the arrivals), but the capacity of the system cannot be measured in an empirical sense.

In a recent paper, Lantz and Rosén (2016) noted this incongruity and suggested that because the relationship between the arrival process, system capacity, and operating characteristics, such as expected queueing time, are theoretically defined within the field of queueing models, it would make more practical sense to use the two objectively measurable entities in the relationship to estimate the third entity that otherwise cannot be estimated objectively. In other words, using an empirical estimate of the arrival intensity and its distribution in combination with the empirically estimated waiting time in the queue, or the empirically estimated length of the queue, can estimate the capacity of the service process without any type of direct observation. This avoids the Hawthorne effect and other types of potential biases. In their study, Lantz and Rosén (2016) used this approach to estimate the effective capacity of the triage process in an emergency department based on detailed empirical data on arrivals and waiting times and without having to study the triage process itself. However, the authors' modelling was restricted to a Markovian arrival and service process, and the authors only briefly discussed the potential of the approach to capacity management.

Therefore, the purpose of this paper is to provide a broader formula-based analysis on the ways that different queueing models can be applied. The analysis is then applied to estimate the capacity of a system in practice and to discuss the general validity and applicability of this approach. A production system, both in manufacturing firms and service organizations, can be designed in several ways (see, e.g. Fitzsimmons and Fitzsimmons 2008; Krajewski et al. 2016). In Cigolini and Grando (2009), two real-life manufacturing systems, a parallel system, and a serial system with tightly interconnected machines, were analysed regarding the system's capacity and productivity. According to the authors, the capacity of a parallel system is estimated by the sum of the capacity of all the machines in the system while the capacity of a serial system is determined by the bottleneck capacity or pace. It is common practice to identify capacity as output that is limited by the bottleneck operation over a period. Determining capacity, therefore, involves defining the bottleneck operation (see, e.g. Boehmer 1982; Gupta 2003), and substantial attention has been focused on the importance (i.e. management) of bottlenecks or the capacity-constraining stage in a production system (see, e.g. Goldratt and Cox 1984; Goldratt 1988 1990; Plenert 1993; Watson et al. 2007). Thus, when we discuss how queueing models can be used to estimate the capacity of a production system, we primarily refer to the system bottleneck but also to other single operations or stages of a system.

The remainder of this paper is organized as follows. In the next section, we discuss the concept of capacity, including methods for capacity measurement. In Section 3, we derive formulas to estimate effective capacity for different types of queueing models. In Section 4, we discuss the practical applications of these formulas. Finally, in Section 5, we present our conclusions on the proposed concept for capacity measurement with recommendations for future research in this field.

2. LITERATURE REVIEW

According to Slack et al. (2010, 299), operations capacity is defined as '*the maximum level of value-added activity over a period of time that the process can achieve under normal operating conditions'*. For instance, capacity can be measured as the maximum number of cars manufactured per day or the maximum number of patients treated per hour. Capacity is, thus, a limitation on output in a process over a specified period (Krajewski et al. 2016).

For operations management, capacity is also defined functionally according to design capacity and effective capacity. A process design capacity is the maximum output that can be achieved under ideal operating conditions, that is, full utilization of productive time, often called '24-7'. Effective capacity is the maximum output a process can achieve under normal conditions (Slack et al. 2010). Design capacity is always greater than effective capacity, and the differences between them can be explained by productive time under normal conditions that is used for setups, maintenance, and breaks. Effective capacity in a process can be calculated as a proportion of the ideal capacity under normal conditions. (Russell and Taylor 2006) Similar concepts, theoretical, and practical capacity, with the same definition as design and effective capacity, are used in the cost accounting literature to estimate the cost (fixed cost allocation) per time or unit space of resources. As a rule of thumb, practical capacity is assumed to be 80−85% of theoretical capacity. (DeBruine and Sopariwala 1994; Hertenstein et al. 2006; Horngren et al. 2012; Kaplan and Anderson 2004, 2007).

In practice, however, defining and measuring capacity under normal operating conditions are not straightforward. Complexity characterizes many operations (Abril et al. 2008; Burdett and Kozan 2006; McNair et al. 2010; Slack et al. 2010). For processes that provide a relatively small number of standardized products or services, capacity in terms of output is relatively easy to define and measure. For processes that provide a more complex mixture of products and services and when the mixture varies over time, measuring capacity in terms of output becomes more complicated. Capacity over a specified period depends on the mix of products or services required during that period, that is, a different mix creates different bottlenecks in the process (Adan and Vissers 2002; Plenert 1993; Slack et al. 2010; Sobreiro 2014; Vastag 2000). Capacity measurement in service organizations is further complicated by the extent of customer participation in the production process (Bamford and Chatziaslan 2009; Corsten and Stuhlmann 1998) and by service providers' experience and skill levels (Combes et al. 2008; Strum et al. 2000). These two conditions increase the variability of service time.

A heterogeneous range of products and services prevents a direct measurement of capacity with respect to output. In this case, it may be more appropriate to define and measure capacity in terms of input, such as rooms in a hotel, seats in a restaurant (Pullman and Rodgers 2010), or time availability such as machine hours and labour hours (Corsten and Stuhlmann 1998; DeBruine and Sopariwala 1994; Kaplan and Anderson 2004; Kaplan et al. 2014; Slack et al. 2010). Doing so assumes that the input and output parameters are proportionally related to each other and demand can be measured by the input units required (Corsten and Stuhlmann 1998). In healthcare, for example, capacity is measured according to resources, such as the total number of beds available, total operating time available per day, and the number of nurses available per day or full-time equivalents (Adan and Vissers 2002; Bamford and Chatziaslan 2009; Kuntz et al. 2007). However, some of these definitions of capacity are questionable because the time dimension is ignored. According to the definition above, the total number of beds in a hospital, total number of rooms in a hotel, or total number of seats in a restaurant do not become a capacity measure until combined with, for example, the number of patients treated per day (Schroeder 1993).

During our review of the literature on the subject, we identified two principal methodological approaches to measuring capacity, measurements based on engineering and physical factors and measurement based on time studies. We elaborate on these approaches below.

2.1 Capacity measurements based on engineering and physical factors

For example, studies that estimate optimum capacity in ports are largely based on engineering and physical factors (Chang et al. 2012; Chu and Huang 2005; Talley 1988, 1994, 2006). According to Chang et al. (2012), the engineering approach to capacity estimation is the current practice when port planners in Asia estimate berth or port capacity for a specified period. The approach can be illustrated by the following formula (Chang et al. 2012, 245).

$$V = N \times C \times E \times K \times H \times D \times O \times U$$

Where

V = annual capacity at a berth measured by the number of containers handled (TEU)
N = number of cranes at berth
C = crane capacity per hour
E = the efficiency ratio of a crane in comparison to its official rate from its design specifications (%)
K = the efficiency after subtracting the interfering effect between cranes (%)
H = number of working hours per day
D = number of working days per year
O = utilization rate of berths (%)
U = the working rate of cranes (%)

V estimates a berth's effective annual capacity and is a measure of a berth's maximal output under certain conditions (i.e. a given level of productive resources). V is thus sensitive to the input values of the variables in the formula and, in some cases, the actual throughput is significantly greater than the estimated berth capacity. This is one of the major weaknesses of the engineering approach to capacity estimation, which also leads to inaccuracies in port capacity planning (Abril et al. 2008; Chang et al. 2012; Chu and Huang 2005; Kozan and Burdett 2005).

Similar methods to estimate design and effective capacity are used in a variety of industries such as airlines (Mirkovic and Tosic 2014; NcNair et al. 2010), airports (Hockaday and Kanafani 1974), breweries (McNair et al. 2015), railways (Abril et al. 2008; Burdett and Kozan 2006; Kozan and Burdett 2005), manufacturing (Grando and Turco 2005), and restaurants (Muller 1999). In the railway industry, for example, the analytical method (i.e. mathematical formulas) is one of the most relevant standard methods to measure theoretical and practical capacity. In this context, theoretical capacity is an approximation used to measure capacity according to the number of trains that could run on the entire railway, on a critical section of rail, or on railway lines during a specific period and under ideal conditions. Practical capacity represents a more realistic measure of rail capacity under normal conditions and is calculated as a proportion of theoretical capacity. Practical capacity is typically around 60–75% of theoretical capacity (Abril et al. 2008; Burdett and Kozan 2006; Kontaxi and Ricci 2012). The capacity in a manufacturing system is defined as the throughput value per unit of time and is normally measured based on factors such as production rate per time unit, plant calendar time (i.e. calendar time minus idle time because of vacations or holy days, for example), net utilization, and defect rate (Grando and Turco 2005; Li et al. 2006; Reid and Bulich 1996).

We draw two conclusions concerning the engineering approach to capacity estimation. First, the method is sensitive to input values of the variables when calculating design (theoretical) capacity. Second, it is difficult to determine what constitutes normal conditions. Thus, it is also difficult to determine the proportional relationship between design and effective (practical) capacity that provides a reliable estimate of effective capacity. This is particularly true in situations where the mixture of products or services produced varies over time.

2.2 Capacity measurement based on time studies

Time studies have been used in manufacturing industries since Frederick W. Taylor created the principles of scientific management in the early 20th century (Taylor 1923). One application of time studies is to determine standard times for production process setup and operations that are primarily used for capacity planning including capacity measurement, operations scheduling, and cost accounting. (see, e.g. Kaplan and Anderson 2004, 2007; Wild 1995).

Time studies are based on direct or indirect time measurements. Direct time studies are conducted by observations, interviews, or surveys while indirect time studies are based on management knowledge and experience, historical data, or work measurement techniques. For work measurement techniques, work tasks are broken down into constituent elements or motions with time standards established for individual motions (Maynard et al. 1948; Razmi and Shakhs-Niyaee 2008; Wild 1995). Work measurement techniques based on standardized time modules are difficult to apply, particularly in service industries where service time varies depending on customer needs and behaviour and on the service providers' skills and ability to provide the requested services. Direct time measurements and indirect time measurements based on management knowledge and experience are, therefore, typically used in service industries. (Bamford and Chatziaslan 2009; Combes et al. 2008; Corsten and Stuhlmann 1998; Larsson 2013; Prajogo 2006; Strum et al. 2000).

In the service industry literature, we found examples of direct and indirect time measurements mainly in the healthcare industry (see, e.g. Finkler et al. 1993; Hollingsworth et al. 1998; Pizziferri et al. 2005; Sittig 1993; Tang et al. 2007; Yen et al. 2009). Therefore, we will continue to focus on time measurements in healthcare. Time-motion studies have been used by hospital managers and researchers since their introduction in the early 20th century to study costs and performance (or inefficiencies) in healthcare processes (Chase and Apte 2007; Lopetegui et al. 2014). According to Lopetegui et al. (2014), three different groups of time-motion data collection methods are used in healthcare; i) external observers who collect time and motion data, ii) self-reporting by active tracking, self-reported work sampling, surveys, interviews, and focus groups, and iii) automatic timestamps, that is, automatic task durations created by computer systems.

Westbury et al. (2009) used surveys to measure the mean duration of several common operations and used the data to estimate the demand for total surgical operation time, that is, needed capacity. According to the authors, this measure of demand might be a more accurate data input for a capacity planning process compared to the usual methods of measuring demand, that is, demand in terms of the number of patients waiting for surgery. Larsson (2013) examined the accuracy of two methods for estimating surgical operations time, one based on estimation by surgeons and the other based on historical data in computer systems. The results of this examination showed that estimations based on computer systems are more accurate than estimations by surgeons. Bratt et al. (1999) used a time-motion study based on direct observations as a benchmark and compared the results with other data collection methods including interviews and self-reports. The results indicated that time data collected by interviews and self-reports largely underestimate non-productive time.

Time-motion data collection methods have been called into question for the behavioural problems that they cause and the measurement errors and inaccurate capacity and cost figures that they can produce. Data collection by interviews, surveys, and self-reports are considered unreliable because the resulting estimates have poor validity. These methods underestimate non-productive time and overestimate productive time significantly because employees have a strong incentive not to report unused time or non-productive time (Balakrishnan et al. 2012; Bratt et al. 1999; Cardinaels and Labro 2008; Kaplan and Anderson 2004, 2007; Lopetegui et al. 2014). A consequence of this unavoidable measurement error is overestimated capacity utilization (Balakrishnan et al. 2012). Data collection by direct observations is questionable because of the potential risk of the Hawthorne effect (Campbell et al. 1995; Franke and Kaul 1979; Lopetegui et al. 2014; Roethlisberger and Dickson 1939; Wickström and Bendix 2000). According to the Hawthorne effect, a process is affected by the fact that it is observed. For example, staff might work either more or less efficiently than they would under normal conditions because they know that their performance is being measured. Moreover, time-motion data collection methods are often considered irritating by employees, which could lead to additional measurement error (Kaplan and Anderson 2007).

In summary, the two principal methodological approaches to capacity measurement identified and discussed in this section suffer from different types of drawbacks. Measurements based on engineering and physical factors are sensitive to the input values of the variables and the proportional relationship between design and effective capacity. Measurements based on time studies are questionable because of the behavioural problems that they can cause, such as the Hawthorne effect and the fact that employees tend to underestimate non-productive time leading to measurement errors. Thus, regardless of which approach to capacity measurement is most suitable for use in a specific situation, measurement bias exists. In the next section, we propose a concept for the measurement of effective capacity based on queueing models, which can overcome these drawbacks.

3. USING QUEUEING MODELS TO ESTIMATE CAPACITY

In a recent paper, Lantz and Rosén (2016) introduced queueing models to obtain effective system capacity estimates. The authors used the theoretical relation between the arrival rate λ, the service rate μ, and the expected queueing time W_q (or length of the queue L_q). Because λ and W_q (or L_q) can typically be estimated objectively based on empirical data, they can be used to derive an objective estimate of μ without direct observation of the service process itself, thereby avoiding, for example, the Hawthorne effect. More formally, the idea was to rearrange the standard queueing formula $W_q = f(\lambda, \mu)$ (or $L_q = f(\lambda, \mu)$) to $\mu = f(\lambda, W_q)$ (or $\mu = f(\lambda, L_q)$) and to use the resulting formula to estimate the capacity in a system based on empirical observations of the average waiting time in the queue (or length of the queue) and arrivals to the system. This is the fundamental idea that this paper develops. Hence, traditional formulas for the expected queueing time and the expected length of the queue are used to derive formulas that can be used to estimate effective capacity in different types of systems based on queue data only. To save space, this paper will only illustrate the idea in detail for an M/M/1 system. However, similar analyses have also been done for the relation between λ, μ, and L_q in M/M/2, M/D/1, M/Er/1, M/G/1, and G/G/1 systems. These results can be found in Lantz and Rosen (2017).

The M/M/1 model is characterized by a Markovian arrival process, exponential service time, and a single server. In other words, the model assumes that customers arrive at the system according to a Poisson distribution and the service time follows an exponential distribution. The mean number of arrivals is λ per unit of time (e.g. per hour), and the mean service time is $1 / \mu$ units of time. Hence, μ is considered effective system capacity in terms of the maximum number of customers that, on average, can be served by the system during one unit of time. The expected waiting time in the queue W_q can be written as

$$(1) \qquad W_q = \frac{\lambda}{\mu(\mu - \lambda)}.$$

Solving (1) for μ yields

$$\mu = \lambda / 2 + \sqrt{(\lambda / 2)^2 + \lambda / W_q} \tag{2}$$

which can be used as a formula to estimate the capacity in an M/M/1 system based on empirical observations of the average waiting time in the queue and the arrivals to the system. For example, if the average waiting time in the queue has been empirically point estimated to be 0.2 hours and the average arrival rate to the system has been empirically point estimated as four jobs per hour, the effective capacity of the system can, by applying equation (2), be point estimated to $4.0/2 + \sqrt{(4.0/2)^2 + 4.0/0.2} = 6.9$ jobs per hour. The expected number of customers in queue L_q in an M/M/1 system can be written as

$$L_q = \frac{\lambda^2}{\mu(\mu - \lambda)}. \tag{3}$$

Solving (3) for μ yields

$$\mu = \frac{\lambda L_q + \lambda \sqrt{L_q^2 + 4L_q}}{2L_q} \tag{4}$$

which can be used as a formula to estimate the capacity in an M/M/1 system based on empirical observations of the average number of customers in the queue and the arrivals to the system. For example, if the average number of customers in the queue of a system has been empirically point estimated to 2.25 customers, and the average arrival rate to the system has been empirically point estimated to three customers per hour, the effective capacity of the system can be point estimated to four customers per hour based on straightforward application of equation (4).

From a theoretical perspective, using point estimated parameters to point estimate another parameter implies that the statistical uncertainty is escalated. In other words, we must consider the combined uncertainty of the point estimated input parameters when using them to create a confidence interval for the true effective capacity. However, it is beyond the scope of this paper to provide a deeper analysis of this phenomenon. Moreover, as Lantz and Rosén (2016) mentioned, a confidence interval for the effective capacity would still be narrow if the analysis is based on large quantities of data.

4. DISCUSSION

Capacity is not easy to measure, particularly in service organizations (McNair et al. 2010; Sasser 1976: Slack et al. 2010). In the literature section of this paper, we identified various deficiencies associated with the two main methodological approaches to measuring effective capacity. The validity of the output of these measurements is questionable because of these deficiencies. However, the suggested methodology to measure effective capacity is based on arrival rates and queueing time (or queue length), which are typically easy to measure objectively with high validity. Queueing models are already applied to capacity planning and control in many industries (see, e.g. Lakshmi and Iyer 2013; Pullman and Rodgers 2010; Rao et al. 1998; Shabayek and Yeung 2000) and could be an appropriate alternative for capacity measurement.

Therefore, how should queueing models be used in practice to measure the capacity of a system based on observed averages for the queueing time (or length of the queue) and the arrival rate? First, the most suitable queueing model for the specific situation must be selected. Parameters such as the number of servers and the queueing discipline are typically known, but the actual distribution of the service process and the arrival process may be more difficult to determine. If jobs are known not to arrive according to a Poisson process to a system with one server, the G/G/1 model should be the natural choice. If jobs do arrive approximately randomly to a system with one server where the variance in the service time is high, obtaining estimates of the variance and then applying the M/G/1 model should be attempted. When jobs arrive approximately randomly to a system with one server where the variance in the service time is low, system capacity lies somewhere between the indications of the M/M/1 model and the M/D/1 model. Typically, the calculated values for μ based on specific values for λ and W_q differ relatively little between the M/M/1 model and the M/D/1 model. Moreover, if data regarding queueing times as well as queue lengths are available, more reliable estimates of capacity can be obtained because the average of these two individual estimates can be used rather than just one.

When jobs arrive approximately randomly to a system with one server where the service process consists of several phases with similar service times in sequence, the M/Ek/1 model is the best option. If the phases are relatively unequal with respect to service time, it is better to view the service process as a whole and use the M/G/1 model to measure its capacity.

For the practical application of the proposed method to measure effective capacity, the model parameters (i.e. λ and W_q or L_q) must be accurately estimated from historical data at an appropriate level of detail (Brown et al. 2005; Dragovic et al. 2006; Gans et al. 2003; Koole and Mandelbaum 2002; Papadopoulos and Heavey 1996). In Lantz and Rosén (2016), the case study was based on a data set extracted from the hospitals admission database, which included the arrival and queueing times for triage for 30,000 arrivals to the emergency department during a one-year period. It was possible to estimate parameter values and to verify the distribution of the arrival process. Similar data, on an individual level, were used in, for example, Brown et al. (2005), Kim et al. (1999) and McManus et al. (2004). In other cases, data are reported or summarized as average in fixed consecutive time periods and not on an individual level. These periodical data are assumed to be Poisson random variables with a constant rate for each period (Brown et al. 2005; Gans et al. 2003; Palvannan and Toew 2012; Shabayek and Yeung 2000). In further cases, arrival data are recorded but not the data on service time (see, e.g. Green et al. 2006 and Liu et al. 2006). When applying the proposed measuring method, a corresponding situation can occur concerning data on waiting times or queue length (Mani et al. 2015). When data are not available, they must be collected, which can be costly (Gans et al. 2003; Green 2006; Palvannan and Teow 2012). A lack of data can limit the application of the proposed capacity measuring method and queuing theory in general (Liu et al. 2006; Lu et al. 2013). However, as information systems are developed and large quantities of data are stored and analyzed at reasonable cost, the availability of relevant data will increase in many industries (Gans et al. 2003; Green et al. 2006; Kesavan and Mani 2015; Lu et al. 2013).

Finally, although parameter assumptions, such as statistical distributions, should be tested when plausible, we believe that it is more important to avoid the potential biases associated with human decision making than to ensure that a certain assumption is 100% valid. As Daniel Kahneman (2011) has shown, human judgement is generally inferior to algorithms. Hence, the ability to measure the capacity of a system in model terms without direct interaction or with direct knowledge of the system is valuable in several respects.

5. CONCLUSION

Accurate capacity data input is essential to effectively fulfil capacity planning in both manufacturing firms and service organizations. It is, therefore, important to develop reliable capacity measurement methods for this purpose. In this study, we have recognized two major methodological approaches to capacity measurement and discussed their shortcomings. To avoid these shortcomings, we have proposed an approach to measure effective capacity indirectly, but objectively, based on queueing models. The estimations are based on two variables, the arrival rate and queueing time (or queue length), which are usually easy to measure objectively and empirically with high validity in contrast to the service rate (i.e. effective capacity). Hence, and this represents the primary argument, the queueing model approach suggested here does not require direct data collection from the shop floor or the application of rules of thumb for the relationship between design and effective capacity. In other words, the conceptual problems related to measurements based on engineering and physical factors as well as those related to time studies can both be avoided since the manufacturing or service process itself is considered a black box when the queueing model approach is applied.

The capacity measurement approach suggested here can be used to estimate the effective capacity of most manufacturing and service processes if the arrival process and the waiting times can be observed and the assumptions of a relevant queueing model are valid.

The fact that queueing models are often used for other purposes in capacity planning and control in a range of industries indicates that the suggested approach to capacity measurement could be an appropriate alternative to traditional methods for capacity measurement in many cases. In a previous study (Lantz and Rosén 2016), we used the queueing model approach to estimate the effective capacity of the triage process in an emergency department of a medium-sized Swedish hospital. During one year, approximately 30,000 patients arrived at the emergency department. The exact time of arrival and the exact waiting time before triage were recorded for each patient on an individual level. Using these data, the expected arrival rate and the expected queueing time during different hours of the day were estimated with some precision, and the capacity of the triage process and its variation during the day was calculated based on the estimates. The results showed several differences in these estimates compared to the traditional way that the capacity of the triage function was measured. First, the effective capacity in the triage process is not a linear function of the number of nurses – the marginal effect on capacity of adding another nurse is decreasing. Second, there is a substantial difference in capacity between an experienced nurse and an inexperienced nurse. Third, actual service levels in the triage process vary substantially during the day and night. These results warrant further research towards the implementation and evaluation of the proposed approach to measuring effective capacity in other types of processes.

Another area of further research is directed toward the development of the proposed measurement methodology. As mentioned earlier, the math quickly becomes complex as the number of servers increases, indicating a need for reliable approximations for more complex queueing models. Approximations will be an unconditional necessity for such complex models where service rate (μ) cannot be derived from the standard queueing formulas.

REFERENCES

Abril, M., F. Barber, L. Ingolotti, M. A. Salido, P. Tormos, and A. Lova. 2008. “An assessment of railway capacity.” *Transportation Research Part E: Logistics and Transportation Review* 44 (5): 774–806.

Adan I. J. B. F., and J. M. H. Vissers. (2002). "Patient mix optimisation in hospital admission planning: a case study." *International Journal of Operations & Production Management* 22 (4): 445-461.

Adenso-Diaz, B., P. González-Torre, and V. García. 2002. “A capacity management model in service industries.” *International Journal of Service Industry Management* 13 (3): 286-302.

Balakrishnan, N., B. Render, and R. M. Stair. 2014. *Managerial Decision Modeling with Spreadsheets*. Pearson Education Ltd, Harlow.

Balakrishnan, R., E. Labro, and K. Sivaramakrishnan. 2012. “Product Costs as Decision Aids: An Analysis of Alternative Approaches (Part 2).” *Accounting Horizons* 26 (1): 21–41.

Bamford, D., and E. Chatziaslan. 2009. “Healthcare capacity measurement.”, *International Journal of Productivity and Performance Management* 58 (8): 748-766.

Boehmer, R. B. 1982. “Capacity: Its Measurement and Management.” In *Handbook of Industrial Engineering,* edited by G. Salvendy, 11.5.1-11.5.13. John Wiley & Sons, New York.

Bratt, J. H., J. Foreit, P. Chen, C. West, B. Janowitz, and T. De Vargas. 1999. “A comparison of four approaches for measuring clinician time use.” *Health Policy and Planning* 14(4): 374–381.

Brown, L., N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, and L. Zhao. 2005. “Statistical Analysis of a Telephone Call Center: A Queueing-Science Perspective.” *Journal of the American Statistical Association* 100 (469): 36-50.

Bruneel, H., D. Fiems, J. Walraevens, and S. Wittevrongel. 2014. “Queueing models for the analysis of communication systems.” *TOP* 22: 421–448.

Burdett, R. L., and E. Kozan. 2006. “Techniques for absolute capacity determination in railways.” *Transportation Research Part B: Methodological* 40 (8): 616-632.

Campbell, J. P., V. A. Maxey, and W.A. Watson. 1995. “Hawthorne Effect: Implications for Prehospital Research.” *Annals of Emergency Medicine* 26 (5): 590-594.

Canonaco, P., P. Legato, R. M. Mazza, and R. Musmanno. 2008. “A queuing network model for the management of berth crane Operations.” *Computers & Operations Research* 35: 2432-2446.

Cardinaels, E., and E. Labro. 2008. “On the Determinants of Measurement Error in Time-Driven Costing.” *The Accounting Review* 83 (3): 735-756.

Chase, R. B., and U. M. Apte. 2007. “A history of research in service operations: What’s the big idea?” *Journal of Operations Management* 25 (2): 375–386.

Chang, Y. T., J. Tongzon, M. Luo, and P. T. W. Lee. 2012. “Estimation of Optimal Handling Capacity of a Container Port: An Economic Approach.” *Transport Reviews* 32 (2): 241–258.

Chu, C., and W. Huang. 2005. “Determining container terminal capacity on the basis of an adopted yard handling system.” *Transport Reviews* 25 (2): 181-199.

Cigolini, R., and A. Grando. 2009. “Modelling capacity and productivity of multi-machine systems.” *Production Planning & Control* 20 (1): 30-39.

Combes, C., N. Meskens, C. Rivat, and J. P. Vandamme. 2008. “Using a KDD process to forecast the duration of surgery.” *International Journal of Production Economics* 112 (1): 279-293.

Corsten, H., and S. Stuhlmann. 1998. “Capacity management in service organisations.” *Technovation* 18 (3): 163–178.

DeBruine, M., and P. R. Sopariwala. 1994. “The Use of Practical Capacity for Better Management Decisions.” *Journal of Cost Management* 8 (1): 25-31.

Defraeye, M., and I. Van Nieuwenhuyse. 2016. “Staffing and scheduling under nonstationary demand for service: A literature review.” *Omega* 58: 4–25.

Dragović, B., N. K. Park, and Z. Radmilović. 2006. “Ship-berth link performance evaluation: simulation and analytical approaches.” *Maritime Policy & Management* 33 (3): 281-299.

Elmaghraby, S. E. 1991. “Manufacturing capacity and its measurement: A critical evaluation.” *Computers & Operations Research* 18 (7): 615-627.

Finkler, S. A., J. R. Knickman, G. Hendrickson, M. Lipkin, and W. G. Thompson. 1993. “A Comparison of Work-Sampling and Time-and-Motion Techniques for Studies in Health Services Research.” *Health Services Research* 28 (5): 577-597.

Fitzsimmons J A, and M. J. Fitzsimmons. 2008. *Service Management: Operations, Strategy, Information Technology*. McGraw-Hill, NY.

Franke, R. H., and J. D. Kaul. 1978. “The Hawthorne Experiments: First Statistical Interpretation.” *American Sociological Review* 43 (5): 623-643.

Gans, N., G. Koole, and A. Mandelbaum. 2003. “Telephone Call Centers: Tutorial, Review, and Research Prospects.” *Manufacturing & Service Operations Management* 5 (2): 79-141.

Goldratt, E. M., and J. Cox. 1984. *The Goal*. North River Press. New York.

Goldratt, E. M. 1988. “Computerized shop floor scheduling.” *International Journal of Production Research* 26 (3): 443–455.

Goldratt, E. M. 1990. *What Is The Thing Called Theory of Constraints, and How Should It Be Implemented?* North River Press. New York.

Govil, M. K., and M. C. Fu. 1999. “Queueing Theory in Manufacturing: A Survey.” *Journal of Manufacturing Systems* 18 (3): 214-240.

Grando, A., and F. Turco. 2005. “Modelling plant capacity and productivity: conceptual framework in a single-machine case.” *Production Planning & Control* 16 (3): 309-322.

Green, L. V. 2006. “Queueing Analysis in Healthcare.” In *Patient Flow: Reducing Delay in Health Care Delivery,* edited by R.W. Hall, 281-307. Springer, NY.

Green, L. V., J. Soares, M. D. Giglio, and, M. D. Green. 2006. “Using Queueing Theory to Increase the Effectiveness of Emergency Department Provider Staffing.” *Academic Emergency Medicine* 13 (1): 61-68.

Gupta, M. 2003. “Contraints management – recent advances and proactices.” *International Journal of Production Research* 41 (4): 647–659.

Hertenstein, J. H., L. Polutnik, and C. J. McNair. 2006. “Capacity Cost Measures and Decisions: Two Field Studies.” *The Journal of Corporate Accounting & Finance* 17 (3): 63-78.

Hockaday, S. L. M., and A. K. Kanafani. 1974. “Developments of Airport Capacity Analysis.” *Transportation Research* 8 (3): 171-180.

Hollingsworth, J. C., C. D. Chisholm, B. K. Giles, W. H. Cordell, and D. R. Nelson. 1998. “How Do Physicians and Nurses Spend Their Time in the Emergency Department?” *Annals of Emergency Medicine* 31 (1): 87-91.

Horngren, C. T., S. M. Datar, and M. V. Rajan. 2012. *Cost Accounting: A Managerial Emphasis*. Pearson Education Ltd, Harlow.

Hwang, J., L. Gao., and W. Jang. 2010. “Joint demand and capacity management in a restaurant system.” *European Journal of Operational Research* 207 (1): 465–472.

Kahneman, D. 2011. *Thinking, fast and slow*. Farrar, Straus and Giroux, New York.

Kaplan, R. S., and S. R. Anderson. 2004. “Time-Driven Activity-Based Costing.” *Harvard Business Review* November: 131–138.

Kaplan, R. S., and S. R. Anderson. 2007. “The Innovation of Time-Driven Activity-Based Costing.” *Journal of Cost Management* 21 (2): 131–138.

Kaplan R. S., M. Witkowski, M. Abbott, A. B. Guzman, L. D. Higgins, J. G. Meara, E. Padden et al. 2014. “Using Time-Driven Activity-Based Costing to Identify Value Improvement Opportunities in Healthcare.” *Journal of Healthcare Management* 59 (6): 399-412.

Karvonen, S., J. Rämö, M. Leijala, and J. Holmström. 2004. “Productivity improvement in heart surgery: a case study on care process development.” *Production Planning & Control* 15 (3): 238-246.

Kesavan, S., and V. Mani. 2015. “An Overview of Industry Practice and Empirical Research in Retail Workforce Management.” In *Retail Supply Chain Management: Quantitative Models and Empirical Studies,* edited by N. Agrawal, and S. A. Smith, 113-145. International Series in Operations Research & Management Science 223, Springer, New York.

Kim S-C., I. Horowitz, K. K. Young, and T. A. Buckley. 1999. “Analysis of capacity management of the intensive care unit in a Hospital.” *European Journal of Operational Research* 115 (1): 36-46.

Klassen, K. J., and T. R. Rohleder. 2001. “Combining Operations and Marketing to Manage Capacity and Demand in Services.” *The Service Industries Journal* 21 (2): 1-30.

Koole, G., and A. Mandelbaum. 2002. “Queueing Models of Call Centers: An Introduction.” *Annals of Operations Research* 113: 41–59.

Kontaxi, E., and S. Ricci. 2012. “Railway Capacity Handbook: A Systematic Approach to Methodologies.” *Procedia - Social and Behavioral Sciences* (48): 2689-2696.

Kozan, E., and R. Burdett. 2005. “A railway capacity determination model and rail access charging methodologies.” *Transportation Planning and Technology* 28 (1): 27-45.

Krajewski, L. J., M. K. Malhotra, and L. P. Ritzman. 2016. *Operations Management: Processes and Supply Chains*. Pearson Education Ltd, Harlow.

Kuntz, L., S. Scholtes, and A. Vera. 2007. “Incorporating efficiency in hospital-capacity planning in Germany.” *European Journal of Health Economics* 8 (3): 213–223.

Lakshmi C., and S. A. Iyer. 2013. “Application of queueing theory in health care: A literature review.” *Operations Research for Health Care* 2 (1-2): 25–39.

Lantz, B. 2015. *Operativ verksamhetsstyrning*. Studentlitteratur (in Swedish).

Lantz, B., and P. Rosén. 2016. "Measuring effective capacity in an emergency department." *Journal of Health Organization and Management* 30 (1): 73-84.

Lantz, B., and P. Rosén. 2017. "Using queueing models to estimate system capacity", *Production Planning & Control* xx (x): xx-xx.

Larsson, A. 2013. “The accuracy of surgery time estimations.” *Production Planning & Control* 24 (10-11): 891-902.

Li, J., D. E. Blumenfeld, and J. M. Alden. 2006. “Comparisons of two-machine line models in throughput analysis.” *International Journal of Production Research* 44 (7): 1375–1398.

Little, J. D. C. 1961. “A Proof for the Queuing Formula: L= λ W.” *Operations Research* 9 (3): 383-387.

Little, J. D. C. 2011. “Little’s Law as Viewed on Its 50th Anniversary.” *Operations Research* 59 (3): 536–549.

Liu, Z., L. Wynter, C. H. Xia, and F. Zhang. 2006. “Parameter inference of queueing models for IT systems using end-to-end measurements.” *Performance Evaluation* 63 (1): 36–60.

Lopetegui M., P. Yen, A. Lai, J. Jeffries, P. Embi, and P. Payne. 2014. “Time motion studies in healthcare: What are we talking about?” *Journal of Biomedical Informatics* 49: 292–299.

Lu, Y., A. Musalem, M. Olivares, and A. Schilkrut. 2013. ”Measuring the Effect of Queues on Customer Purchases.” *Management Science* 59 (8): 1743-1763.

Mani, V., S. Kesavan, and J. M. Swaminathan. 2015. “Estimating the Impact of Understaffing on Sales and Profitability in Retail Stores.” *Production and Operations Management* 24 (2): 201–218.

Maynard H. B., G. J. Stegemerten, and J. L. Schwab. 1948. *Methods-Time Measurement*. MacGraw-Hill, NY.

McManus, M. L., M. C. Long, A. Cooper, and E. Litvak. 2004. “Queuing Theory Accurately Models the Need for Critical Care Resources.” *Anesthesiology* 100 (5): 1271-1276.

McNair, C. J., L. Polutnik, H. H. Johnston, J. Augustyn, and C. R. Thomas. 2003. “Shifting Perspective: Accounting. Visibility, and Management Action.” In *Advances in Management Accounting 11*, edited by M. J. Epstein, and M. A. Malina, 1-39.

McNair, C. J., T. Watts, V. Baard, and L. Polutnik. 2010. “Improving productive potential in the airline industry by exploring the productive limits of capacity.” *International Journal Critical Accounting* 2 (4): 372-398.

Mirkovic, B., and V. Tosic. 2014. “Airport apron capacity: estimation, representation, and flexibility.” *Journal of Advanced Transportation* 48 (2): 97–118.

Ng, I. C. L., J. Wirtz, and K. S. Lee. 1999. “The strategic role of unused service capacity.” *International Journal of Service Industry Management* 10 (2): 211-244.

Muller, C. C. 1999. “A simple measure of restaurant efficiency.” *The Cornell Hotel and Restaurant Administration Quarterly* 40 (3): 31-37.

Palvannan, R. K., and K. L. Teow. 2012. “Queueing for Healthcare.” *Journal of medical systems* 36: 541–547.

Papadopoulos, H. T., and C. Heavey. 1996. “Queueing theory in manufacturing systems analysis and design: A classification of models for production and transfer lines.” *European Journal of Operational Research* 92 (1): 1-27.

Pizziferria, L., A. F. Kittlera, L. A. Volka, M. M. Honourb, S. Guptaa, S. Wanga, T. Wanga et al. 2005. “Primary care physician time utilization before and after implementation of an electronic health record: A time-motion study.” *Journal of Biomedical Informatics* 38 (3): 176–188.

Plenert, G. 1993. “Optimizing theory of constraints when multiple constrained resources exist.” *European Journal of Operational Research* 70 (1): 126-133.

Prajogo, D. 2006. “The implementation of operations management techniques in service organisations: An Australian perspective.” *International Journal of Operations & Production Management* 26 (12): 1374-1390.

Pullman, M., and S. Rodgers. 2010. “Capacity management for hospitality and tourism: A review of current approaches.” *International Journal of Hospitality Management* 29 (1): 177–187.

Razmi, J., and M. Shakhs-Niyaee. 2008. “Developing a specific predetermined time study approach: an empirical study in a car industry.” *Production Planning & Control* 19 (5): 454-460.

Rao, S. S., A. Gunasekaran, S. K. Goyal, and T. Martikainen. (1998). “Waiting line model applications in manufacturing.” *International Journal of Production Economics* 54 (1): 1-28.

Reid, R. A., and M. Bulich. 1996. “Traditional and quantitative modeling approaches in production capacity analysis.” *Production and Inventory Management Journal* 37 (2): 21-25.

Roethlisberger, F. J., and W. Dickson. 1939. *Management of the Worker*. Harvard University Press, Cambridge, MA.

Russell, R. S., and B. W. Taylor. 2006. *Operations Management: Quality and Competitiveness in a Global Environment*. John Wiley & Sons, NJ.

Sasser, W. E. 1976. “Match supply and demand in service industries.” *Harvard Business Review* 54 (6): 133-140.

Schroeder R. G. 1993. *Operations Management: Decision Making in the Operations Function*. McGraw-Hill, New York, NY.

Shabayek, A. A., and W. W. Yeung. 2000. “A queuing model analysis of the performance of the Hong Kong container terminals.” *Transportation Planning and Technology* 23 (4): 323-351.

Shemwell Jr, D. J., and J. Cronin. 1994. "Services Marketing Strategies for Coping with Demand/Supply Imbalances." *Journal of Services Marketing* 8 (4): 14-24.

Sittig, D. F. 1993. “Work-Sampling: A Statistical Approach to evaluation of the Effect of Computers on Work Patterns in Healthcare.” *Methods of Information in Medicine* 32 (2): 167-174.

Slack, N, S. Chambers, and R. Johnston. 2010. *Operations Management*. Pearson Education Ltd, Harlow.

Sobreiro, V. A., E. B. Mariano, and M. S. Nagano. 2014. “Product mix: the approach of throughput per day.” *Production Planning & Control* 25 (12): 1015-1027.

Stahlbock, R., and S. Voß. 2008. “Operations research at container terminals: a literature update.” *OR Spectrum* 30: 1–52.

Strum, D. P., A. R. Sampson, J. H. May, and L. G. Vargas. 2000. “Surgeon and Type of Anesthesia Predict Variability in Surgical Procedure Times.” *Anesthesiology* 92 (5): 1454–66.

Taha, H. A. 1981. “Queueing Theory in Practice.” *Interfaces* 11 (1): 43-49.

Talley, W. K. 1988. “Optimum throughput and performance evaluation of marine terminals.” *Maritime Policy & Management* 15 (4): 327-331.

Talley, W. K. 1994. “Performance indicators and port performance evaluation.” *Logistics and Transportation Review* 30 (4): 339-352.

Talley, W. K. 2006. “Chapter 22 Port Performance: An Economics Perspective.” In *Devolution, Port Governance and Port Performance*, edited by M. R. Brooks, and K. Cullinane, 499-516. Research in Transportation Economics 17. Elsevier Ltd, Oxford.

Tang, Z., L. Weavind, J. Mazabob, E. J. Thomas, M. Y. L. Chu-Weininger, and T. R. Johnson. 2007. “Workflow in intensive care unit remote monitoring: A time-and-motion study.” *Critical care medicine* 35 (9): 2057 -2063.

Taylor, F. W. 1923. *The principles of scientific management*. Harper, New York.

Vastag, G. 2000. “The theory of performance frontiers.” *Journal of Operations Management* 18 (3): 353–360.

Watson, K. J., J. H. Blackstone, and S. C. Gardiner. 2007. “The evolution of a management philosophy: The Theory of constraints.” *Journal of Operations Management* 25 (2): 387–402.

Westbury, S., M. Pandit, and J. J. Pandit. 2009. "Matching surgical operating capacity to demand using estimates of operating times." *Journal of Health Organization and Management* 23 (5): 554-567.

Wickström, G., and T. Bendix. 2000. “The “Hawthorne effect” – what did the original Hawthorne studies actually shows?” *Scandinavian Journal of Work, Environmental & Health* 26 (4): 363-367.

Wild, R. 1995. *Production and Operations Management*. Cassel Educational Ltd, London.

Yen, K., E. L. Shane, S. S. Pawar, N. D. Schwendel, R. J. Zimmanck, and M. H. Gorelick. 2009. “Time Motion Study in a Pediatric Emergency Department Before and After Computer Physician Order Entry.” *Annals of Emergency Medicine* 53 (4): 462-469.