*3rd International Conference on Industrial Engineering and Industrial Management*
*XIII Congreso de Ingeniería de Organización*
*Barcelona-Terrassa, September 2nd-4th 2009*

# SERVICE NEEDS FORECASTING: An Approach for the Motor Industry Using Analogies with Medical ER Management Models

**Miguel Ángel Perez Salaverría**

## Abstract

*During the past years, the industry has shifted position and moved towards "the luxury universe" whose customers are demanding, treating individuals as unique and valued customer for the business, offering vehicles produced with the state of the art technologies and implementing the highest finishing standards. Due to the competitive level in the market, motor makers enable processes which equalizes customer services to E.R. management, being dealt with the maximum urgency that allows the comparison between both, car workshops and emergency rooms, where workshop bays or ramps will be equal to emergency boxes and skilled technicians are equivalent to the health care specialist, who will carry out tests and checks prior to afford any final operation, keeping the "patient" under control before it is back to normal utilization.*

*This paper ratify a valid model for the automotive industry to estimate customer service demand forecasting under variable demand conditions using analogies with patient demand models used for the medical ER.*

**Keywords:** Service level - Variable demand – Luxury universe – Service cost

## 1. Introduction

Motor makers experience product-related requests from customers that are used to align the product with market's necessities improving the attractiveness of the product. However, these requests could become disruptive when including threats that unless the feature is added, changed, fixed etc., the customer will not buy the product, will stop servicing it or sell it.

As a rule, car manufacturers build strong service networks, but, even in the best case, a gap still remains between customer real demands and Authorized Service retailer's workload. Formerly, premium brands were being focused on reaching high service standards to match their customer expectations on service and maintenance, placing price on a secondary option, while volume car maker's acts basically on pricing and service times. According to premium brands point of view, a car entering the workshop is treated as a matter of urgency; it is like a patient entering an emergency room of a hospital and needs to be diagnosed with regards to the symptoms present in this moment, to offer the best solution for this particular case.

The seminal references found for the present document are based on the works of B. Liu, who, in 1976, established an input-output approach for regional hospital needs projection. Later, in 1982, R. D. Kamenetzky, L. J. Shuman and H. Wolfe studied a how to estimate necessities and demands for prehospital care.

Usually, motor makers make their facilities capacity estimations using empirical methods considering various restrictions, but typically customers arrivals occurs under uncertain

conditions and variable demand that were not included in the calculations, producing work overloads, stocks backorders and customer complaints. Opposite to that, medical emergency services apply different techniques to dimension hospital facilities according to the demographic distribution of the area to be serviced. Estimations are compared with a computer model simulation result and validated to create a model to be applied in future health services.

Other authors afford the case from an operational research point of view, such as: A. Bagust, M. Place and J. W. Posnett studied in 1999 a dynamic model to be used for accommodating emergency admissions applying stochastic simulation. Later in 2004, S. C. Brailsford, V. A. Lattimer, P. Tarnaras and J. C. S. Turnbull studied an emergency and on-demand health care model for large complex systems. Then in 2006, L. V. Green, S. Savin and B. Wang studied a model to manage patient service in a diagnostic medical facility.

Subsequently in 1989, D. M. Rhyne reviewed the applicability and a measure of forecasting systems in managing hospital services demand. In 1993, M. A. Badri and J. Hollingsworth published a simulation model for scheduling in the emergency room. Also, in 1996, Y. Gerchak, D. Gupta and M. studied a reservation planning under uncertain demand for emergency surgery.

This paper substantiate the process to accommodate the existing models used for medical facilities to the service needs of a car service network and explores experimental procedures used in ER management for comparing the capabilities of complex discrete event service systems. Instead of measuring system capability by analyzing or simulating the system with a constant rate of arriving work, system capability is measured as the maximum rate of work arrival for which the system has a steady state. Hence, we seek the arrival rate which causes the system to be at full capacity. This rate is arguably the best indication of the service system's capability.

## 2. A general service model

Inclusion of more ad more electronic devices interacting together in the car makes requires a better understanding of vehicle electrical architecture and has an impact on training needs, modern facilities with nice and clean workshops and, of course, a good management to ensure the required productivity and efficiency. Opposite to that, generally, low salaries still offered to the workshop technicians enabling a high personnel rotation.

The former statement supposes any skilled technician will act as experienced doctor to diagnose a critical patient in an emergency box. The service receptionist will therefore assign jobs and times to the workshop according to pre-established priorities rivalling the medical ER. Customer requests can be a double-edged sword. On the one hand they can help point the way of where the market wants a company to go. On the other, requests can become disruptive and distracting. By understating the factors behind customer requests, the dynamics of the relationship and how these requests impact the process, companies can channel the "request energy" into positive channels leading to a better product that customers are excited about and willing to pay for.

A maxim of the analysis of service systems is the structure will have stationary long-run behaviour if and only if the number of arriving tasks are, on average, less than the number of tasks the system is capable to process.

The service systems considered are centralized, controllable and do not generate tasks at a rate A per unit time. Tasks are admitted upon generation and processed by the system; completed tasks are ejected from the system that has the capability to process as many tasks per unit time on average. If our overall system can work at a maximum of p tasks per unit time, we can input as many as p per unit time and the system will remain stationary. If A is our arrival rate for the system, we wish to manipulate A to expose p.

Work-conserving queuing models do not allow tasks to expire or to create other tasks while in service. Those tasks will not split or combine and always finish service. Work-conserving queuing system models are common in both the practice and literature of applied probability. In a typical experiment, we generate input to the system at a constant rate, monitor the performance of the system either at fixed intervals or upon departure from the system, and employ well known methods of steady-state analysis to estimate the steady-state average of the performance measure.

The service level constraint formulation allows for expansion policies that either anticipates demand reaching the capacity position or react to demand having exceeded it. Its evaluation by using barrier option pricing tools is exact, and therefore the numerical results supersede those where timing and size decisions were made sequentially and evaluation of the service level constraint could err on the side of caution. The optimal expansion parameters nearly always increased or decreased together. The delayed and infrequent expansion strategy that corresponds to large values of both parameters is optimal when greater shortages are permissible, lead times are short, economies of scale are significant, average demand growth is small, and/or demand volatility is low.

The opposite strategy, of small and frequent expansions that are initiated proactively, is optimal when the problem parameters reacts a more stringent service level, smaller economies of scale, and greater risk of shortage from the combination of long lead times and faster or more volatile demand growth.

Lastly, a deterministic lead time was considered for expansion. A probability distribution could be considered for lead times to make it more realistic and the act of stochastic lead time on the capacity expansion problem could be analyzed.

## 3. Capacity reserve

Few estimations of hospital cost structures have taken account of this aspect of hospital production and none have been applied. Freidman and Pauly (1983) and Gaynor and Anderson (1995), have all incorporated the impact of stochastic demand on hospital cost structures, while also recognising that hospitals control the output decisions, in response to such demand. In these studies the emphasis has been on estimating the cost of maintaining reserve capacity.

### 3.1. Full capacity

Hospitals reserve capacity in response to demand uncertainty to aid the specification for optimal capacity, which incorporates reserve capacity costs.

Running at full capacity also imposes a cost, however, in the form of production inflexibility, leading to patients being queued or turned away. There is therefore a trade-off between the

cost of holding unused capacity in order to service stochastic demand, and operating at full capacity and turning patients away.

This trade-off defines the optimal level of reserve capacity compatible with economically efficient utilisation. As Gaynor and Vogt (2000) note in any case, failure to take account of stochastic demand and the consequent production responses, leads to misspecification of hospital cost-output relations.

### 3.2. Optimal capacity

The resolution of optimal capacity depends on the appropriate specification for outputs. One limitation of previous studies is they have used aggregate measures of hospital total admissions to define output.

A second limitation of previous studies is the reliance on annual or quarterly fluctuations in demand to model hospital responses to stochastic demand. It seems more realistic to model shorter-term fluctuations in demand to capture such responses.

Use of aggregate measures, for both hospital output and demand fluctuation, will lead to a loss of information on the form and structure of the demand uncertainty. The precise stochastic nature of demand will vary according to the type of case being serviced. There are two ways to generate data from a work-conserving system which will reveal the maximum processing rate in the system. They are:

- input tasks to the system at a rate known to be much higher than the system can handle

- fill the system, then input a new task every time that a task completes

In the former, the rate of outgoing jobs eventually converges to p. Instead of choosing a very high input rate and dealing with the problems of exploding buffer contents and a no recurrent system, we will simply close off the system and recalculate the tasks which finish. Hence, we take the second approach.

### 3.3. Elective and stochastic demand admissions

Hospitals distinguish between elective and emergency admissions. Each hospital allocate the fixed capacity based on their expectations of emergency demand turning into effective demand, recognising that these expectations may not be realised ex post.

Demand for emergency services is assumed randomly distributed with a known probability density function, while there is an assumed excess demand for elective treatments. Hospital referrals are designated to be emergency or elective cases with waiting lists used to explicitly ration the capacity allocated to elective treatments.

Simultaneously each individual hospital retains some capacity to meet stochastic emergency demand, while also maintaining a waiting list for elective demand. In order to produce at any given level of output the hospital commits resources ex ante based on a forecast of emergency demand. Given seasonal fluctuations and the short-term nature of hospital planning such forecasts are based on within-year variations, even although budget allocations are tied to a yearly cycle.

## 4. Forecasting with limited data using ARIMA models

A time series is a set of observations ordered according to the time they were observed. As the value observed at time t may depend on values observed at previous time points, time series data may invalidate independence assumptions.

An ARIMA(p, d, q) model can be used for temporal dependence in several ways. First, the time series is differenced to render it stationary, by taking d differences. Second, the time dependence of the stationary process is modelled by including p auto-regressive and q moving-average terms, in addition to any time-varying covariates. For a cyclical time series, these steps can be repeated according to the period of the cycle, whether quarterly or monthly or another time interval. ARIMA models are extremely flexible for continuous data.

It should be noted that not all choices of parameters produce well-behaved models. In particular, if the model is required to be stationary then conditions on these parameters must be met.

### 4.1. Forecasting hospital demand using ARIMA models

Hospital bed managers face a difficult task in attempting to allocate their beds between emergency admissions and so-called elective admissions, which are planned and, in general, referred by the patient's doctors or consultants. Depleting the bed availability in an attempt to clear waiting lists runs the risk of being unable to admit emergency cases. On the other hand a policy of reserving too many beds for emergency admissions has an obvious impact on waiting lists.

The conceptual motivation for the empirical variable cost model estimated below follows that of Freidman and Pauly (1983) and Gaynor and Anderson (1995). Following the latter a short-run cost model is estimated with attention focussed on how hospitals use existing fixed capacity to service unexpected demand. Variable hospital costs are specified as a function of the in-patient output, disaggregated into emergency and elective outputs, as well as other dimensions of output such as day case, accident and emergency and outpatient activity, and other characteristics of the hospital such as teaching status. An estimate of the level of fixed resource use, measuring the extent of excess capacity is incorporated through the inverse occupancy rate, which also controls for length of stay.

All these cost elements are conditioned on the hospital’s estimate of unexpected demand as it relates to the probability of the hospital being full. This is controlled for through an estimate of unexpected emergency demand that enables empirical testing of whether or not uncertainty impacts hospital costs. It is hypothesised that if the coefficient on this variable is positive and significant, then demand uncertainty imposes a real cost on hospital production. It is this variable that differentiates the approach from the traditional cost function.

The small number of studies which have estimated such a variable have used different estimates of demand uncertainty as proxies for the standby capacity required to service unexpected demand. Gaynor and Anderson (1995) use the first two moments of the distribution of annual demand to proxy the relationship between unexpected demand and standby capacity. Of course the annual level of data smoothes within period fluctuations while the focus on the described distribution emphasises the predictive content of the information used. Freidman and Pauly (1983) employ a measure of the ratio of expected to actual demand analysed on a quarterly basis.

Given that a ratio is estimated, the level of uncertainty is not captured. Indeed such measures of demand uncertainty reflect the expected fluctuations in demand, i.e. the ones the hospitals can predict. If hospitals do accurately predict the fluctuations then there is no reason to expect this to impact on costs. In the model estimated below demand uncertainty is based on a residual estimate of forecast monthly emergency demand. The level of uncertainty faced by a hospital is thus defined as the difference between realised and forecast emergency demand. Such a measure captures the shocks imposed by stochastic excess demand while simultaneously avoiding possible co linearity between these demands.

Following Freidman and Pauly (1983), a simple autoregressive process was modelled assuming demand expectations are related to prior demand experience. Panel data were used to estimate the demand-forecast equation for emergency admissions, and the performance criteria rest on their ability to forecast, rather than explain behavioural relationships.

In the short-run, while the overall capacity is fixed, there is still a choice over the level of different outputs. Maintaining consistency with the theoretical specification, beds are separated into those allocated to the elective sector and those to the emergency sector. These are calculated on the basis that, under conditions of excess demand, occupancy rate in the elective sector is assumed to be 100%, which is consistent with the existence of substantial hospital waiting lists for elective treatments. The level of staffed elective beds is therefore based on elective admissions and length of stay in that sector.

The remaining service availability is assumed to be used for urgent admissions, including an element of reserve capacity. This enables the staffed beds allocated to each sector and the level of reserve capacity in the emergency sector to be determined.

There is no theoretically accepted functional form for hospital cost functions consequently to determine an appropriate functional form a Box–Cox transformation applied to both dependent and explanatory variables was initially estimated. The results suggested a square root transformation on the dependent variable would fit the data with reasonable

Aletras et al. (1997) review the literature with regards to economies of scale finding that economies of scale are exploited at a relatively low level. As they point out, however, this conclusion is premised on the assumption that hospitals are operating on their efficiency frontier. Gaynor and Vogt (2000) note such conclusions are based on inconsistent estimates, and scale economies have to be related to demand uncertainty and production responses.

As an alternative specification a transcendental logarithmic (translog) function was also estimated, but the results (which again can be obtained from the authors) were poor, with counterintuitive signs on the coefficients and insignificant t-statistics on almost all the independent variables.

The marginal costs of emergency and elective admissions are based on the variable cost element, which is taken from the estimated coefficient on the admission variables in the cost equation, and the quasi-fixed element taken from the beds variables. The quasi-fixed element is adjusted for length of stay in the emergency and elective sectors, respectively.

## 5. Service model development

There are some approaches to this type of problem. In terms of the way in which data was gathered over time it seemed perfectly natural to treat the problem as one of times series prediction.

The initial proposition is select a dealer in a local area to enable visits on a weekly basis to check the model development and future updates. Data will be collected from dealer management system (DMS) used by brand franchised workshops to control operational productivity.

A second data collection will be downloaded from the brand warranty management systems to compare stochastic demand and expected visits. With this data comparison we will be able to understand both kinds distribution and study the particularities of the temporary component of the distribution for a given brand.

One of the reports supplied by the brand shows non expected visits in different categories by model, vehicle system, repair process or number of visits. In order to limit the study to a suitable dimension, the selected service should comply with the following conditions:

- Enabling the study of two premium brands.
- Ease the data collection from a management software pack.
- Offer similar customers typology and characteristics.
- Have similar facilities avoiding seasonal differences and other external factors.
- Possibility to avoid management and productivity factors among both brands services.

After data is collected and treated we are able to classify the information by breakdown typology (model, frequency, systems affected, cost, jobs arranged Vs stochastic demand) or workshop average benefit (spare parts sales, % first pick availability)

As result of this classification, data can be processed to be shorted by model. The new datasheet will be processed following the steps below:

- 1st data period analysis (Statistics & Forecasts)
- Simulation model definition
- Data simulation and corrective coefficients definition
- 2nd data period forecast
- Forecast results analysis in comparison with the real values
- FINAL MODELING

## 6. Data analysis

Once the data was collect from the Dealer Management System, both samples can be processed using any of the statistical tools available in the market. In this particular case, the

software used was a free license program called XLSTAT, which can be used as a Microsoft Excel add in.

### 6.1. Considerations to the model

Breakdowns and maintenance tasks occurs randomly in time, but there is a seasonal nature component during summer and other holiday periods and previous weeks which can generate a system input peak to the organization management. First, because every customer is willing to have his vehicle fixed and maintained, but also due to the higher mileage for old cars and less skilled drivers.

### 6.2. 1st data sample analysis

In the case of study, it is noticed the total demand per month is not increasing on a yearly basis. Yearly variations seems to be higher and monthly variations are cyclical each next year.

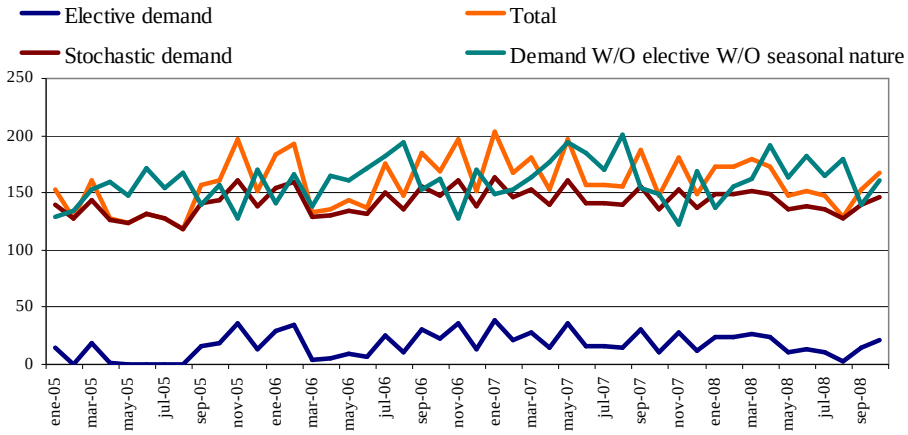


**Figure 1** – System inputs (demand) tender analysis

In order to get the yearly variations clear, we take the Neperian logarithm of the decomposed series, without seasonal nature. A new graphic is obtained showing less noticeable yearly variations.

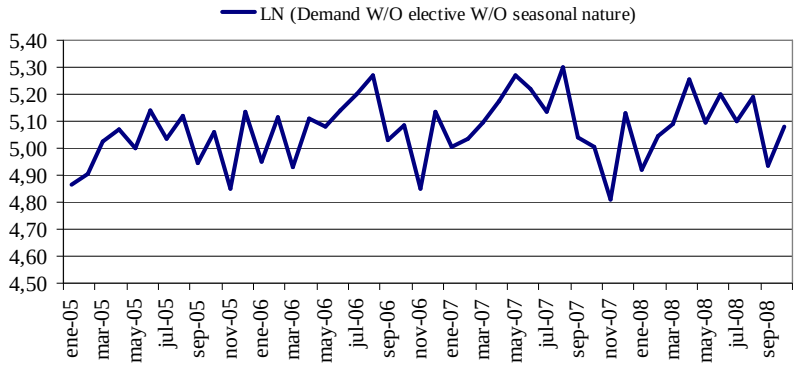


**Figure 2** – LN (Demand W/O elective W/O seasonal nature)

Once the data has been treated, a complete statistical analysis can be obtained by running the different tools in XLSTAT. Selecting the option XLSTAT-SWINDLE in the toolbar and the

command XLSTAT/XLSTAT-Swindle, an ARIMA report can be carried together with a complete analysis of the data.

**Table 1.** XLSTAT – ARIMA report

Centring: YES
Parameters: p = 0 / d = 1 / q = 0 / P = 0 / D = 1 / Q = 1 / s = 46
Optimice: Verosimilitude (Convergence = 0.00001 / Iterations = 500)
Intervals of confidence (%) : 95

Descriptive statistics:

| Variable | Minimum |
|---|---|
| 129.444.306.930.693 | 122,851 |
| **Observations** | **Maximun** |
| 45 | 200,366 |
| **Obs. With lost data** | **Average** |
| 0 | 160,839 |
| **Obs. Without data** | **Stantard Deviation** |
| 45 | 18,357 |

The "Series to analyze" corresponds to the series studied, the data LN (Demand without seasonal nature without prior appointment). The option is left "to Centre" activated in order to permit XLSTAT centring the series automatically.

**Table 2.** XLSTAT – ARIMA White noise test

| Statistic | GDL | Value | p-value |
|---|---|---|---|
| Jarque-Bera | 2 | 0,450 | 0,799 |
| Box-Pierce | 6 | 20,885 | 0,002 |
| Ljung-Box | 6 | 23,842 | 0,001 |
| McLeod-Li | 6 | 22,791 | 0,001 |
| Box-Pierce | 12 | 44,275 | < 0,0001 |
| Ljung-Box | 12 | 55,536 | < 0,0001 |
| McLeod-Li | 12 | 54,086 | < 0,0001 |

### 6.3. 2nd data sample forecast

With the ARIMA information shown on Table 1, a model can be formulated for each sample of data and future periods could be then foretold. In the given example, the final formula was compared to the last 6 months real values to confirm the results were appropriate and use them to make former adjustments. The final equation will look like the following:

$$X(t+1) = Y(t+1)+X(t-1)+X(t-n)-X(t-m) \quad (1)$$

## 7. Conclusions

The results suggest that services do incur costs in holding reserve capacity to service stochastic demand. By separating out this stochastic demand from the excess elective demand

it is possible to quantify this cost. If brand regulatory policies are to be guided by analysis of service costs such considerations are of paramount importance. The setting of labour fees and service levels depends on the accurate demand forecasting, cost of service and understanding of their influence. In turn, fees should be set at a level that provides the appropriate incentives to workshops to hold reserve capacity where this is an efficient response to demand uncertainty.

In this application the various measures of marginal cost and scale economies seemed plausible and consistent with our conceptual arguments relating to production responses to demand uncertainty. Therefore, the data used allows a more detailed specification of hospital output can be applied to the automotive service industry to forecast service requisites.

Furthermore, apparent inefficiencies resulting from services operating within production possibility frontiers may be explained by the existence of uncertain demand, therefore, care should be taken in the interpretation of efficiency rankings without adequate adjustment for demand uncertainty and its impact on cost structures.

Automotive industry is very related to their customer's requirements reaching occasionally slavishness provided to assure future sales and its continuity in the market. This situation grazes the operating limits at present, since clients do not accept a negative answer, services and manufacturers must afford costs when they have not been able to reserve sufficient capacity to attend these demands. Likewise, if they are mistaken upon reckoning a high reserve of capacity they will incur in expenses if occupation is lower than expected.

This paper has evidenced the advantageousness of using ARIMA models similarity to ER to forecast motor industry service demands levels.

**References**

Andrews, K. (2000). Factors that affect the demand for medical care services: A micro-macro econometric analysis. (Ph.D., Clemson University).

Beraldi, P., Bruni, M. E., & Conforti, D. (2004). Designing robust emergency medical service via stochastic programming. European Journal of Operational Research, 158(1), 183.

Brailsford S.C., Lattimer V.A., Tarnaras P. and Turnbull J.C., Emergency and On-Demand Health Care: Modelling a Large Complex System", Journal of the Operational Research Society, 55, 2004, pp 34-42.

Champion, R., Kinsman, L. D., Lee, G. A., & Masman, K. A. (2007). Forecasting emergency department presentations. Australian Health Review, 31(1), 83.

Dawson, D., Jacobs, R., Martin, S., & Smith, P. (2006). The impact of patient choice and waiting time on the demand for health care: Results from the London patient choice project. Applied Economics, 38(12), 5.

Finarelli, H. J.,Jr, & Johnson, T. (2004). Effective demand forecasting in 9 steps. Healthcare Financial Management, 58(11), 52.

Gaur, V., Kesavan, S., Raman, A., & Fisher, M. L. (2007). Estimating demand uncertainty using judgmental forecasts. Manufacturing & Service Operations Management, 9(4), 480.

Gaynor M., Anderson G. F. (1995). Uncertain Demand, The Structure of Hospital Costs, and the Cost of Empty Hospital Beds. NBER Working Papers 4460, National Bureau of Economic Research, Inc.

Gaynor, Martin S. and Vogt, William B., Competition Among Hospitals(November 26, 2002). Available at SSRN: http://ssrn.com/abstract=350920 or DOI: 10.2139/ssrn.350920

Hughes, D., & McGuire, A. (2003). Stochastic demand, production responses and hospital costs. Journal of Health Economics, 22(6), 999.

Jones, S. A., Joy, M. P., & Pearson, J. (2002). Forecasting demand of emergency care. Health care management science, 5(4), 297.

Liu, B. (1976). Regional hospital needs projection: An input-output approach. Socio-Economic Planning Sciences, 10(1), 37-42.

Melnick, G. A., Nawathe, A. C., Bamezai, A., & Green, L. (2004). Emergency department capacity and access in California, 1990-2001: An economic analysis. Health affairs, , W136.

Perkins, W. J. (1996). An analysis of the demand for emergency medical service in a medium sized city. (M.P.H., New York Medical College).

Tempelmeier, H. (2007). On the stochastic uncapacitated dynamic single-item lot sizing problem with service level constraints. European Journal of Operational Research, 181(1), 184.

Turnbul S. C., Brailsford V. A., Lattimer P. and Tarnaras, J. C. (2004). Emergency and on-demand health care: Modelling a large complex system. The Journal of the Operational Research Society, 55(1), 34.

Wang, B. (2003). Capacity management in stochastic service systems. (Ph.D., Columbia University).

Wang, Y., Cohen, M. A., & Zheng, Y. (2002). Differentiating customer service on the basis of delivery lead-times. IIE Transactions, 34(11), 979.