Demand Forecasting and Capacity Management for Hospitals

Article · September 2011

2 authors, including:

Oscar Barros
University of Chile
48 PUBLICATIONS **225** CITATIONS

SEE PROFILE

DEMAND FORECASTING AND CAPACITY MANAGEMENT FOR HOSPITALS

Oscar Barros[1], Richard Weber, Carlos Reveco, Eduardo Ferro, and Cristian Julio

Department of Industrial Engineering, University of Chile, Santiago, Chile

Abstract:

Demand forecasting and capacity management are complicated tasks for certain healthcare services due to the inherent uncertainty, complex relationships, and typically high public exposure involved. Health service demand in three Chilean hospitals has been studied concluding that it can be forecast with high accuracy using Neural Networks and Support Vector Regression. This investigation has allowed us design a process to manage demand by transforming the respective demand forecasts into the resources needed to proper attention. Comparing required resources with available resources and simulating various scenarios permits taking corrective actions when capacity is not aligned with demand.

The proposed forecasting methods and the capacity management process have been accepted by hospital management and staff and are currently in use in one hospital. To support the efficient use of the developed forecasting and management methods, advanced IT systems have been implemented that allow the routine use of the respective processes. We are currently implementing processes and systems in one of the other participating hospitals. The results have been so encouraging that National Health Authorities are considering the extension of the proposed demand forecasting and management practices to close to one hundred public hospitals in Chile.

Keywords: Health care management, forecasting, process management, capacity management

[1] Corresponding author: obarros@dii.uchile.cl; Republica 701, Santiago, Chile, tel. 562 9784037, fax 562 9784011

1. INTRODUCTION

Public hospitals in Chile have, in general, more demand for health services than available capacity permits to attend within a reasonable time frame. Hence it is important for a hospital to forecast demand with great precision, in order to adjust capacity or take alternative courses of action, *e.g*. transferring patients to other facilities. For example, it is possible to discharge patients from a hospital to a local health service center for non-complex pathologies; also private services can be hired in case of an emergency that cannot be treated at a public hospital. Since demand forecasts are not sufficient on their own, but serve as input for hospital management, it is required that service demand be predicted not only on an aggregated level but for different pathology types separately, which makes them technically more demanding.

The forecasted demand for each pathology type allows determining the required resources, such as doctors of different specialties, reception areas, emergency room cubicle capacity and operating room capacity. Comparing the resources needed to satisfy demand with available capacity permits making decisions to adjust capacity or prevent or transfer demand.

Public hospitals in Chile, which process 75% of the country's demand for health services (MIDEPLAN, 2006) are not using any formal way of forecasting demand and managing capacity. Current procedures are informal and defined based on the experience of the participants in the process; furthermore such procedures are mainly oriented to solving the problem of excess demand when it occurs. To be fair, there are some informal attempts to foresee how bad the winter period, when most excess demand is produced, is going to be and to make some decisions regarding the number of doctors and hospital beds that will be made available during this season at a given hospital.

Given the situation outlined, we agreed with the Chilean Health Authority to perform an applied research program that would use state-of-the-art analytical tools, process design methodologies, and IT to develop a general solution for demand forecasting and capacity management that could eventually be used at all Chilean hospitals.

Benefits expected from this work are:

- Significantly improved service to hospital patients, with a shorter waiting time.
- Better use of resources for the entire health system, due to a better distribution of demand to the level that could best serve it.
- Better use of resources at each hospital, since their planning can be made with advanced knowledge regarding demand that allows capacity optimization.

We started the research in March, 2009 and selected three hospitals to be studied to develop the methods, processes, and systems that will eventually be used in all Chilean hospitals.

Demand forecasting and management is part of a larger design that intends to provide a systematic solution to global hospital management. Such a solution is based on the design of a general process structure that we developed for hospitals and that defines the management processes which are needed to ensure a predefined service level for patients and to optimize the use of the required resources. The general process structure allowed us to determine the key processes where implementation of new practices would generate most value (Barros and Julio, 2010, 2011). In agreement with health authorities we selected the process described here and another one related to operating room scheduling. In each of the selected hospitals we evaluated

the current situation of demand forecasting and capacity management to determine the feasibility of introducing analytical and formal practices to improve the respective processes.

The results we present in this paper have been developed in collaboration with hospital staff. They reviewed each of the steps described below, which led to a working process for forecasting and managing demand. Emphasis is also given to the experiences we documented during this work and that could be beneficial in similar future projects.

Section 2 of this paper reviews the literature on the use of analytical methods in forecasting and other experiences in hospital capacity management. Section 3 presents how hospital demand has been modeled using several methods with the results obtained. The processes that convert forecasts into the resources needed to satisfy demand and manage capacity are described in Section 4. Section 5 presents conclusions and provides suggestions for future work.

2. REVIEW OF RELEVANT EXPERIENCE

Demand forecasting is a useful and well-studied subject (Armstrong, 2001) that has generated important results in different areas, such as the retail industry (Aburto and Weber, 2007) and inventory control at several enterprises such as Dell (Kapuscinki *et a*l, 2004). Forecasts provide relevant information for making decisions on the resources needed to provide adequate service to meet the potential demand and to avoid stock breakdowns or overstocking.

There is another line of demand forecasts focused on services. In it the variable to predict is the number of clients who will demand the service, in order to manage capacity needed to provide a given level of service. In a recent work, joint demand and capacity management have been proposed for services in a restaurant (Hwang *et al*, 2010) where the main focus lies on optimizing revenue for a given dynamic demand without considering, however, demand forecasting explicitly. A similar study has been proposed for scheduling elective surgery under uncertainty (Min and Yih, 2010) but again without considering uncertain demand which is the main focus of our paper.

In the case of hospital services the capacity is determined by available physical facilities, such as medical cubicles, operating rooms and beds, and human resources, such as doctors, who perform diagnostic procedures and treatments on patients. This capacity should be planned to guarantee a given service level and optimize use of resources; for this an accurate forecast of the number and type of patients who will arrive in the future is needed.

Many different methods have been proposed for forecasting (Armstrong, 2001; Box *et al*, 1994), and several studies compare such methods in terms of accuracy of results. One of these studies that is relevant to our work compares Neural Networks with other econometric methods and concludes that the former give, in general, better results (Adya and Collopy, 1998). As will be shown below, in our experiments the technique of Support Vector Regression can even outperform Neural Networks.

Few studies of formal demand forecast in the health area have been published. Some of these have focused mostly on predicting the number of beds required to meet emergency demand (Jones *et al*, 2002; Schweigler *et a*l, 2009; Farmer and Emani, 1990). These studies have focused on forecasting demand in the emergency room where all patients must be attended to, even with a considerable delay. This is important because there is no possibility of changing the appointment to another date, or of having patients leave without attention, which is relevant to

the input data, because historical demand is equal to the number of patients attended. This fact will be important for the present work, since we were only able to find good data for emergency services. Several studies have shown, however, that in practice a small difference between patient arrival and care service could exist (Kennedy *et al*, 2008) a fact that has been taken care of in our system (see Section 4.4). Another work that uses an approach similar to ours is reported in Shirxia *et al*, (2009) but we will show that our approach provides superior results. For capacity management the usual procedure has been to simulate the flow of patients through emergency facilities. None of the papers we have reviewed considers an explicit state of the art demand forecasting technique, except the one by Marmor *et al*, (2009) that estimates demand based on a long term moving average over the demand. Other papers that use the common approach of static arrival distribution are the following: Garcia *et al*, (1995); Samaha *et al*, (2003); Rojas and Garavito, (2008); and Khurma and Bacioiu, (2008).

3. FORECASTING METHODS: APPLICATIONS AND RESULTS

In this section we describe exploratory analysis and preprocessing the available data, show how the different forecasting models has been generated and present the respective results.

3.1. Analysis and Pre-processing of Available Data for Forecasting

To be able to forecast effectively one of the key factors is the quality of historical information. In addition, the hospital operating conditions and environment should remain relatively stable. This work focuses on two public pediatric hospitals: Luis Calvo Mackenna (from now on referred to as HLCM) and Exequiel González Cortés (HEGC), and a general purpose hospital, San Borja Arriarán (HSBA). These hospitals have quality data in their emergency room areas.

However, to turn this quality data into useful information for the forecasting models, further analyses and a series of transformations were necessary. By analyzing the demand that arrives at the emergency department outliers were detected. We found that two months had substantially higher demand than the average of the remainder of the months and decided therefore to replace them with the respective average; see Figure 1. In both cases of removed outliers one particular situation occurred: a special kind of flu virus led to the abnormally increased demand which consequently could not be predicted based on historical demand data. As a solution we propose to use expert knowledge of the physicians with experience in emergency room data in order to adjust the forecast proposed by our system in cases where such special events happen.

Visual inspection of aggregated demand shown in Figure 1 reveals a strong seasonal pattern. We observe a low demand during the summer months (December-January-February) and a high influx of patients during the months of the winter season (May – June – July; in the southern hemisphere). This is due to the fact that high air pollution, smog, and low temperatures lead to respiratory diseases increasing the number of emergencies. In general a downward trend can be observed over the years.

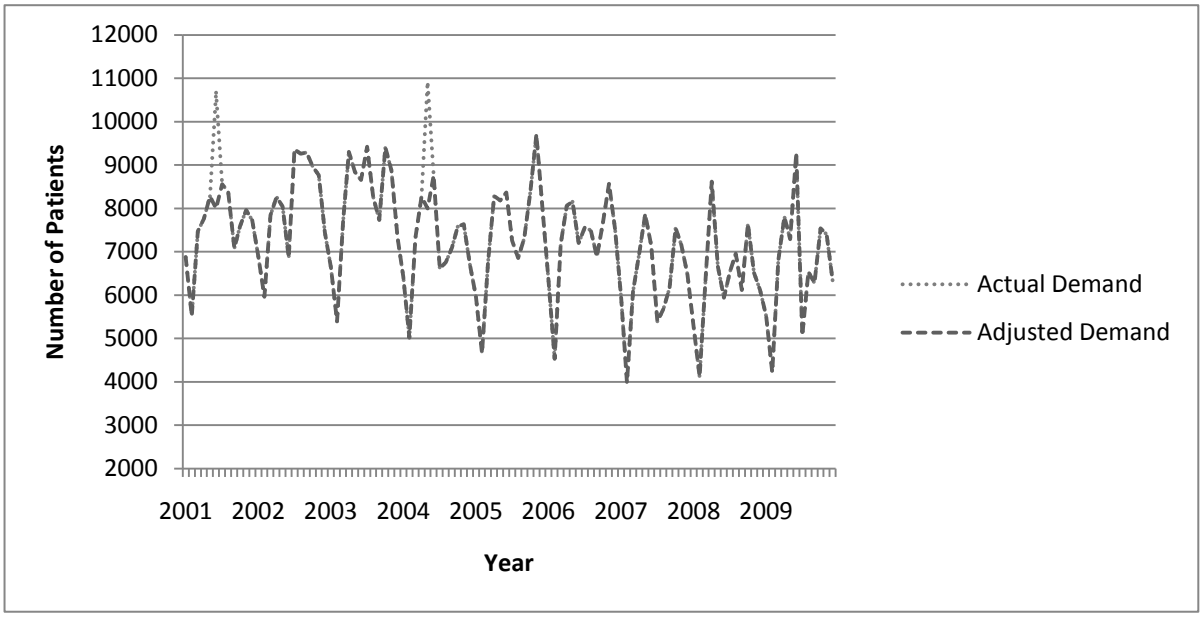

Figure 1: Actual vs. Adjusted Demand per Month in Luis Calvo Mackenna Hospital

When data is disaggregated by pathology type, *e.g*. medical and surgical, we notice huge differences: the first one is much more volatile since it depends on factors such as temperature and influenza like illness rate, as suggested in Jones *et al*, (2002) while the second one is more stable, as shown in Figures 2 and 3. From the data it is also possible to conclude that medical demand comprises 70% of the emergency cases and surgical demand corresponds to 30% of the cases.

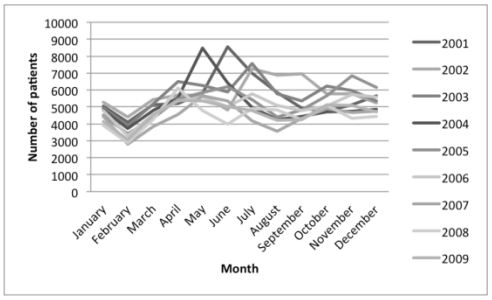

Figure 2: Medical demand for HLCM

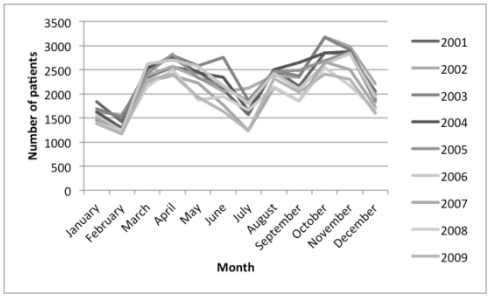

Figure 3: Surgical demand for HLCM

Demands at HEGC and HLCM have a very similar behavior, since both are children hospitals with comparable size and target populations. HSBA data also follows a similar pattern.

On arrival at the emergency facilities each patient is registered, including personal data, time of arrival, diagnosis, and classification according to severity of illness. For the purposes of this work we obtained all this historical data for the three hospitals as follows:

- HLCM: from January, 2001 to December, 2009
- HEGC: from January, 2001 to July, 2009
- HSBA: from January, 2000 to December, 2009.

For the purpose of capacity management it would also be interesting to have the exact time when medical attention starts. This could differ significantly from arrival time but it has not been registered to date.

Unusual demand for treatment of pathologies that appear occasionally, such as allergies and A H1N1, was also discarded, because there was not enough data to detect a pattern; and, in general, outliers were discarded replacing them with an average as mentioned earlier. Daily individual data for patients was aggregated for each month to conform to the time series that we modeled.

This data cleaning procedure was also performed to HEGC and HSBA data.

3.2. Forecasting Methods and their Testing

Four forecasting methods were tested: Linear Regression, Weighted Moving Averages, Neural Networks, as suggested in McLaughlin and Hays, (2008), and Support Vector Regression (SVR). The first two are well known techniques used for forecasting and described in the literature (see Armstrong, 2001). Neural Networks and SVR are recently used techniques for forecasting and will be summarized below.

The particular type of network we used is the Multi-layer Perceptron (MLP). Its basic units are neurons that are grouped in layers and are connected by means of weighted links between two layers. Each neuron receives inputs from other neurons and generates a result that depends only on the information locally available and which serves as input to other neurons. The architecture of the network is shown in Figure 4.

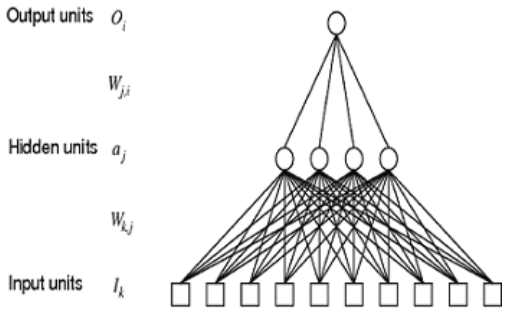

Figure 4: Architecture of a Neural Network (MLP).

Each neuron operates according to the structure in Figure 5, where the output y is determined as a function of the weighted inputs.

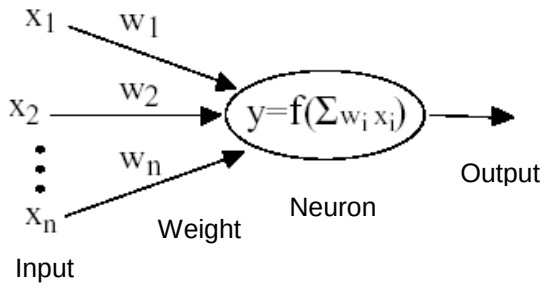

Figure 5: Neuron Details

Function f in Figure 5 is called the activation function and may take different forms; the most commonly used one for continuous outputs is the logistic function, as shown in Eq. (1).

$$f(x) = \frac{1}{1 + e^{-x}} \tag{1}$$

The network was trained with the above mentioned historical data. The basic idea is that previous data predicts a given future month. In particular we assumed that the pattern was seasonal and therefore used previous values of the month we want to predict as some of the inputs to the model. The structure of the network consists of an output layer with one neuron that generates the desired forecast. The input layer contains the variables we will use to explain the demand. In the hidden layer we used a number of neurons between input and output neurons, since a high number will tend to copy the data (over fitting) and a small number will not produce good forecasts.

As mentioned already, previous months were used as input data. However, there are months that are more relevant than others. We tried to select relevant attributes using a genetic algorithm, as suggested in Shirxia et al, (2009), but results were not encouraging. In Shirxia et al, (2009) a common pitfall in neural network design was made, which is to separate the data set into just two groups: one for training and one for testing (Zhang, 2007). This results in trying to minimize the error over the testing data and leads to an over fitting of the resulting model. In our case, we divided the data into three sets: 70% for training; 20% for testing, where the network is trained to minimize the test error. The third set with 10% of the data is independently used to validate results. This use of an independent set provides a better evaluation of future results.

Regarding the network architecture we tested several parameters, such as the number of epochs to use, the learning rate, and the number of hidden neurons. Best results were obtained for 10,000 training epochs, maintaining the model with minimum error in the training set; a learning rate of 0.2 with a momentum of 0.3. Also, decaying was introduced, but this only helps to get to the solution faster with no significant changes in results.

Based on results which will be shown later, we selected a Neural Network with 18 input neurons. If N is the index of the month to be forecast , three neurons corresponding to the values of the same month in previous years, N-12, N-24, and N-36, were included; 3 neurons representing the tendency between months given by the differences between N-12 and N-13, N-

24 and N-25, N-36 and N-37 and a set of 12 binary variables to represent the months of the year are also part of the network`s input layer. This provided a solution that can forecast up to a year in advance and takes account of tendencies. Thus the network has 18 input neurons plus an additional bias neuron that helps to separate cases and allows having smaller Neural Networks than without this bias.

The output layer contains simply one neuron that generates the forecasted demand in month N. The hidden layer contains 10 neurons providing the model an adequate degree of freedom, usually calculated by (Number of input neurons + Number of output neurons)/2. The resulting network is shown in Figure 6.

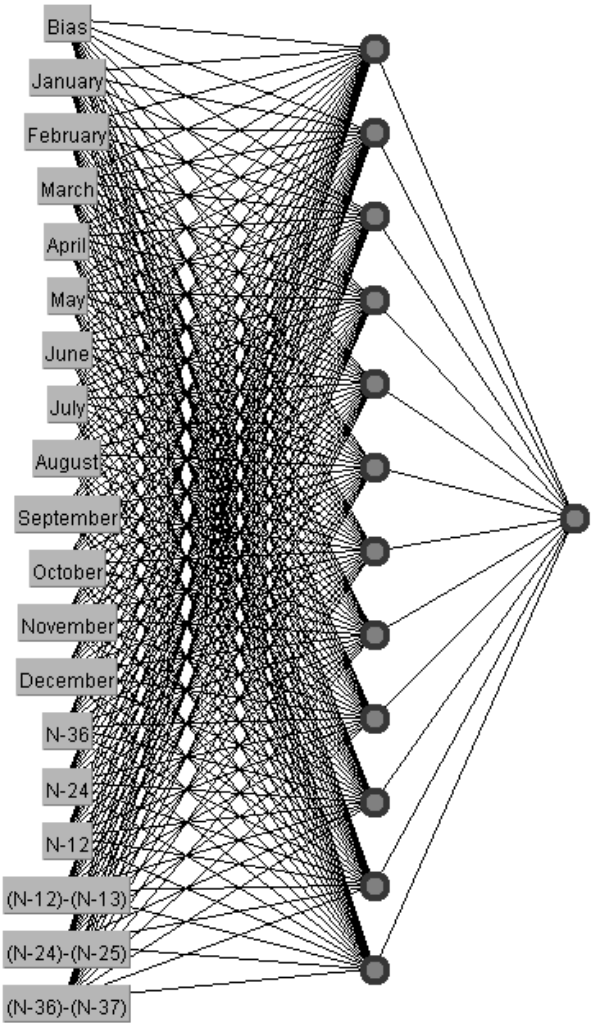

Figure 6: Resulting Neural Network Architecture

Another method we tested is Support Vector Regression (Chen and Schölkopf, 2005; Hofmann *et al*, 2008; Smola and Schölkopf, 2004), which is a variation of Vector Support Machines (SVM) based on the following idea. SV regression (SVR) performs linear regression in

a high-dimensional feature space generated by a kernel function as described below, using the ε-insensitive loss function proposed by Vapnik, (1995). This function allows a tolerance degree to errors not greater than ε as shown in Figure 7. The description is based on the structure and terminology used in Smola and Schölkopf, (2004).

We start with a set of training data $\{(x_1, y_1), \ldots, (x_l, y_l)\}$, where each $x_i \in \mathbb{R}^n$ denotes the input space of the sample and has a corresponding target value $y_i \in \mathbb{R}$ for $i = 1, \ldots, l$, l being the number of available data points to build the regression model. The SVR algorithm applies a function Φ transforming the original data points from the initial Input Space ($\mathbb{R}^n$) to a generally higher dimensional feature space ($F \subset \mathbb{R}^m$). In this new space, a linear model f is constructed, which represents a non-linear model in the original space:

$$\Phi: \mathbb{R}^n \rightarrow F \tag{2}$$

$$f(x) = \langle \omega, \Phi(x) \rangle + b \qquad \text{With } \omega \in \mathbb{R}^m \text{ and } b \in \mathbb{R} \tag{3}$$

In Eq. (3), $\langle \cdot, \cdot \rangle$ denotes the dot product in $\mathbb{R}^m$. When the identity function is used, i.e. $\Phi(x) \rightarrow x$, no transformation is carried out and linear SVR models are obtained.

The goal when using the ε-insensitive loss function is to find a function f that fits given training data with a deviation less or equal to ε and, at the same time, is as flat as possible in order to reduce model complexity. This means that one seeks a small weight vector ω. One way to ensure this is by minimizing the norm $\|\omega\|^2$ (Smola and Schölkopf, 2004), leading to the following optimization problem:

$$\min_{\omega,b} \frac{1}{2} \|\omega\|^2 \tag{4}$$

$$s.t. \begin{cases} y_i - \langle \omega, \Phi(x) \rangle - b \leq \varepsilon \\ \langle \omega, \Phi(x) \rangle - y_i + b \leq \varepsilon \end{cases} \tag{5}$$

This problem could be infeasible. Therefore, slack variables ξ_i, ξ_i^* $i = 1, \ldots, l$, are introduced to allow error levels greater than ε (see Figure 7), arriving at the formulation in Eqs. (6) to (9).

$$\min_{\omega,b} \frac{1}{2} \|\omega\|^2 + C \sum_{i=1}^{l} (\xi_i + \xi_i^*) \tag{6}$$

s.t.

$$y_i - \langle \omega, \Phi(x) \rangle - b \leq \varepsilon + \xi_i^* \tag{7}$$

$$\langle \omega, \Phi(x) \rangle - y_i + b \leq \varepsilon + \xi_i^* \tag{8}$$

$$\xi_i^*, \ \xi_i \geq 0 \tag{9}$$

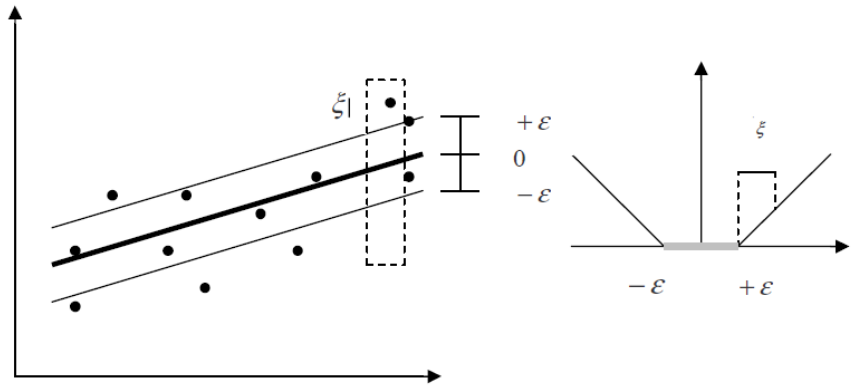

Figure 7: Support Vector Regression to Fit a Tube with Radius ε to the Data and Positive Slack Variables ξi

This is known as the primal problem of the SVR algorithm. The objective function takes into account generalization ability and accuracy in the training set, and embodies the structural risk minimization principle (Vapnik, 1998). Parameter C > 0 determines the trade-off between generalization ability and accuracy in the training data, and the value up to which deviations larger than ε are tolerated. The ε -intensive loss function $|\xi|_\varepsilon$ has been defined as in Eq. 10.

$$|\xi|_\varepsilon = \begin{cases} 0 & if\,|\xi| < \varepsilon \\ |\xi| - \varepsilon & otherwise \end{cases} \tag{10}$$

It is more convenient to represent this optimization problem in its dual form. For this purpose, a Lagrange function is constructed and, once applying saddle point conditions, the dual problem in Eqs. (11) to (14) is obtained (Vapnik, 1995).

$$\max_{\alpha,\alpha^*} -\frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\langle\Phi(x_i),\Phi(x_j)\rangle - \varepsilon\sum_{i=1}^{l}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{l} y_i(\alpha_i - \alpha_i^*) \tag{11}$$

s.t.

$$\alpha_i, \alpha_i^* \le C \quad \forall\, i = 1, \dots, l \tag{12}$$

$$\alpha_i, \alpha_i^* \ge 0 \quad \forall\, i = 1, \dots, l \tag{13}$$

$$\textstyle\sum_{i=1}^{l}(\alpha_i - \alpha_i^*) = 0 \tag{14}$$

This is the quadratic optimization problem that has to be solved to obtain the solution of the SVR model, which is a function of the dual variables α_i and α_i^*. Using saddle point conditions it can be shown that Eq. (15) holds (Vapnik, 1998). Replacing this expression in Eq. (3), the final solution of the SVR algorithm is obtained as Eq. (16).

$$\omega = \sum_{i=1}^{l} (\alpha_i - \alpha_i^*) \Phi(x_i) \tag{15}$$

$$f(x) = \sum_{i=1}^{l} (\alpha_i - \alpha_i^*) K(x_i, x) + b \tag{16}$$

Here, the expression $K(x_i, x)$ is equal to $\langle \Phi(x_i), \Phi(x) \rangle$, which is known as the Kernel Function (Vapnik, 1995). The existence of such a function allows us to obtain a solution for the original regression problem, without knowing explicitly the transformation $\Phi(x)$ applied to the data. Some examples of common Kernel Functions are shown in Table 1.

Name	Kernel Functions
Linear	$x_i \cdot x_j$
Polynomial	$\left[(x_i \cdot x_j) + 1\right]^d$
Radio Basis Function	$\exp\left\{-\gamma \left| x_i - x_j \right|^2\right\}$

Table 1: Common Kernel Functions

It is well-known that the generalization performance of SVR (estimation accuracy) depends on a good setting of parameters C, ε and the kernel parameters (Vladimir, Ma, 2004). We use a Grid-Search to find good parameters for SVR with Radial Basis Function as kernel (γ) within previously specified ranges for parameters C, ε, and γ. In our case we use, $C=10^{-1}, 10^{0} \ldots 10^{10}$; $\varepsilon=2^{-10}, 2^{-9}, \ldots, 2^{-1}$; $\gamma = 2^{-8}, 2^{-7}, \ldots, 2^{0}$.

3.3. Results

We used Mean Average Percentage Error (MAPE) and Mean Square Error (MSE) as performance measures to determine model accuracy, as defined in Eqs. (17) and (18).

$$MAPE = \frac{1}{n} \sum_{i=1}^{n} \frac{|X_i - X_i^{true}|}{X_i^{true}} \tag{17}$$

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (X_i - X_i^{true})^2 \tag{18}$$

$$X_i : Forecasted\ Demand$$

$$X_i^{true}: Actual\ Demand$$

For the Linear Regression, Weighted Moving Average, and SVR the same inputs as the ones described for the Neural Network were used (except for the bias). Results obtained using these four methods for the validation sets of all hospitals are displayed in Table 2.

	Linear Regression		**Weighted Moving Average**		**Neural Network**		**SVR**	
	MAPE	**MSE**	**MAPE**	**MSE**	**MAPE**	**MSE**	**MAPE**	**MSE**
HLCM Medical Demand	12.67%	150,686	7.53%	**144,729**	7.45%	161,689	**5.61%**	154,861
HLCM Surgery Demand	6.54%	27,097	7.36%	**20,137**	8.99%	22,947	**5.09%**	25,199
HEGC Medical Demand	15.91%	3,114,376	16.5%	1,978,332	7.7%	1,043,753	**6.86%**	**606,324**
HEGC Surgery Demand	8.55%	14,302	8.96%	11,730	8.3%	12,155	**5.88%**	**8,120**
HEGC Orthopedic Surgery Demand	8.41%	35,940	8.60%	28,247	5.12%	29,851	**4.44%**	**25,460**
HSBA Medical Demand	8.27%	3,125,071	11.83 %	850,342	7.9%	1,226,165	**6.97%**	**643,984**
HSBA Maternity Demand	10.54%	23,738	6.98%	12,408	10.6%	38,629	**3.24%**	**7,867**

Table 2: Forecast Result Errors (best results in bold)

As shown in Table 2, in four out of seven cases best results are obtained with SVR, when using MSE as criterion to compare the performance of the different models. When using MAPE as criterion for comparison, SVR appears as the best option for demand forecasting in all cases. These results were obtained using Rapid Miner 4.6.000, the Neural Network library from WEKA and SVR from LIBSVM (Chang and Lin, 2001) Library but using Rapid Miner as graphic user interface (GUI).

Figures 8 and 9 display the results graphically showing forecast with SVR and actual demand for HLCM, with a 90% confidence interval for the forecast. This interval has been calculated based on the hypothesis that the forecast error has a normal distribution with a mean zero. This hypothesis has been confirmed using a Kolmogorov-Smirnov test.

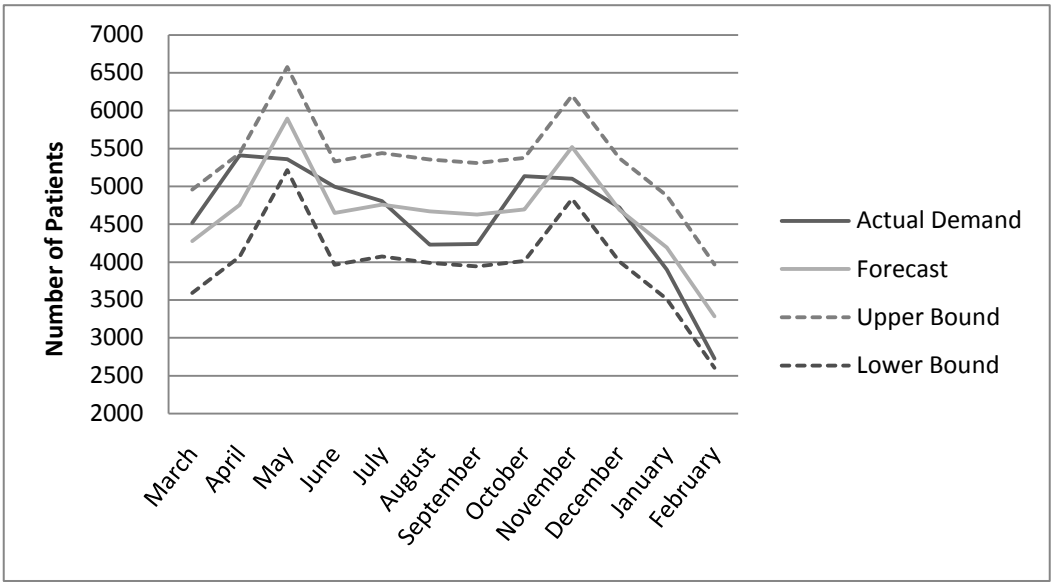

Figure 8: Medical Demand in Luis Calvo Mackenna Hospital

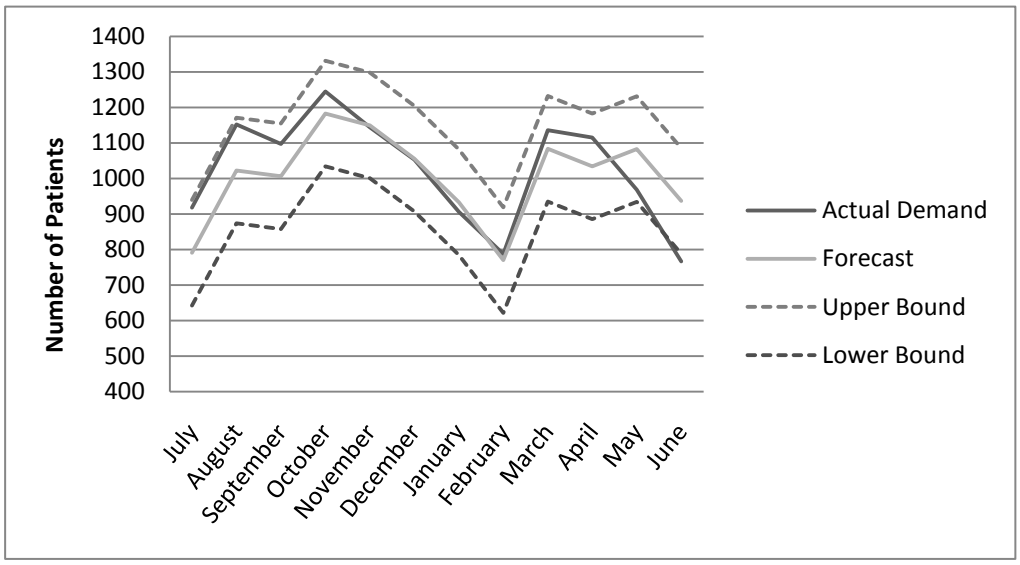

Figure 9: Surgical Demand in Luis Calvo Mackenna Hospital

Similar results are shown in Figures 10, 11 and 12 for HEGC, and in Figures 13 and 14 for HSBA.

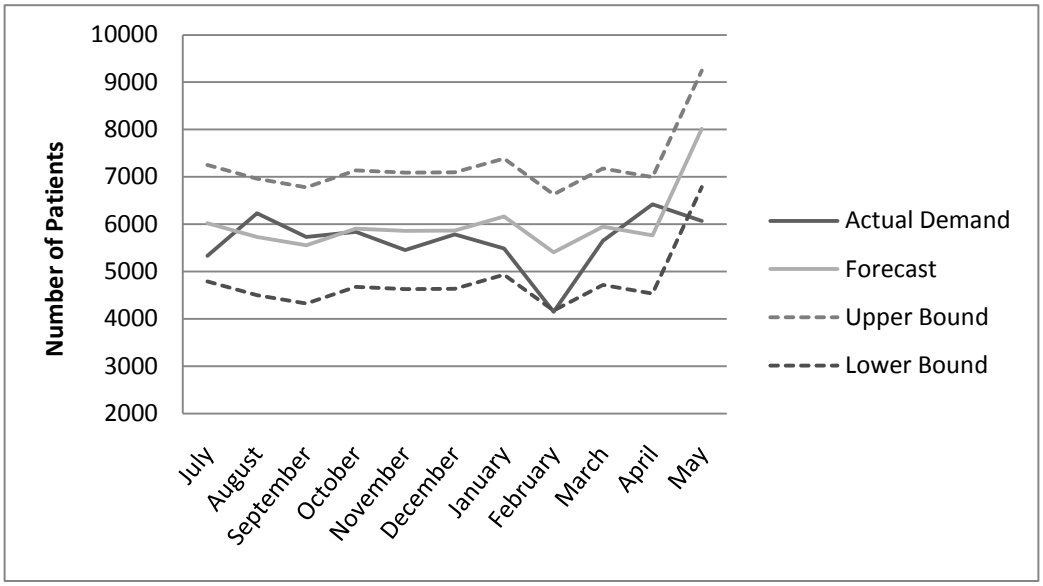

Figure 10: Medical Demand in Exequiel Gonzales Cortes Hospital

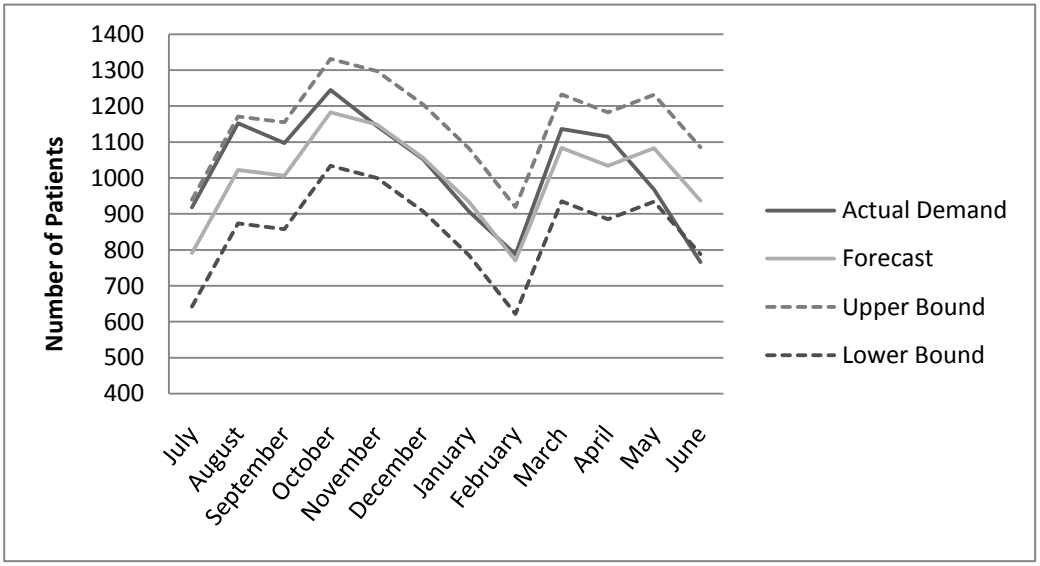

Figure 11: Surgical Demand in Exequiel Gonzales Cortes Hospital

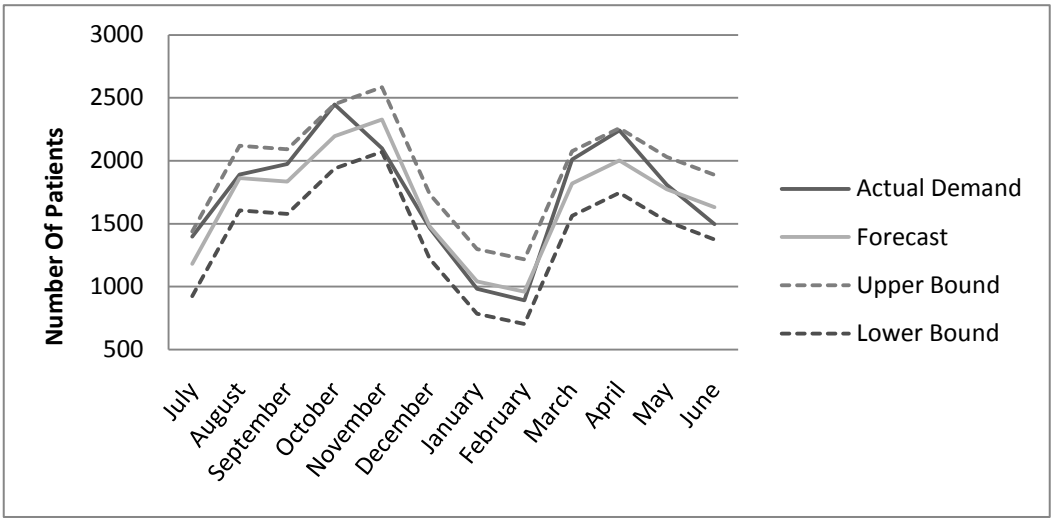

Figure 12: Orthopaedic Surgical Demand in Exequiel Gonzales Cortes Hospital

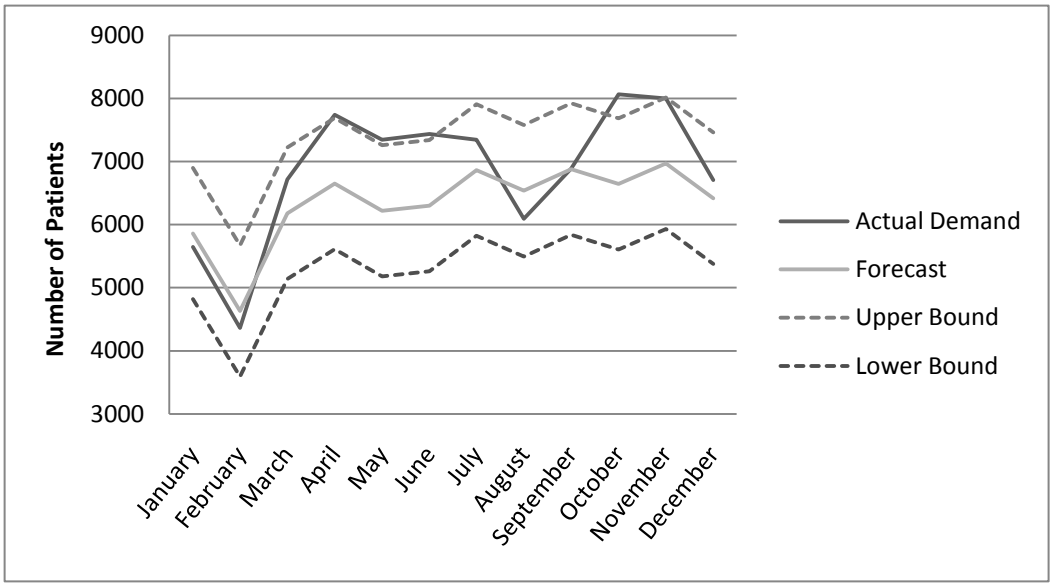

Figure 13: Medical Demand in San Borja Arriarán Hospital

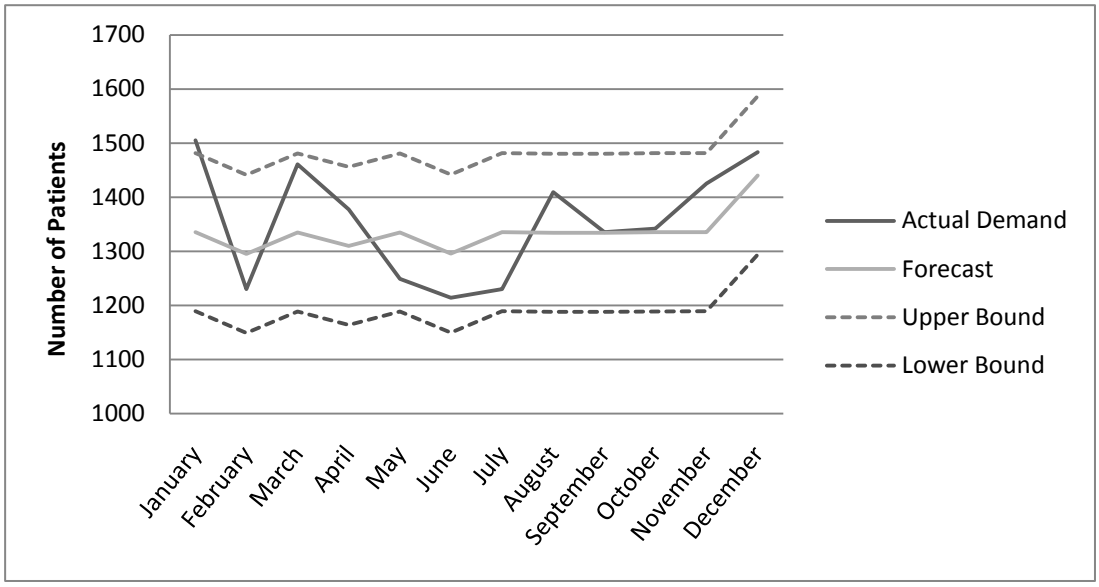

Figure 14: Maternity Demand in San Borja Arriarán Hospital

Based on the results presented above, we conclude that Support Vector Regression is an appropriate method for predicting demand in hospitals, but without discarding the possibility of using simpler methods that may provide acceptable results under certain conditions. Since all the aforementioned models generate forecasts in less than one minute on a standard PC, the run time is irrelevant for the proposed monthly use f the algorithm.

4. CAPACITY MANAGEMENT BASED ON THE FORECAST

As stated in (Jones et al, 2002), having a forecast is not, by itself, a useful contribution to hospital management. Managers also need to know whether their capacity will suffice to attend such demand with defined quality standards, and how the hospital resources could be re-arranged or modified to achieve that goal. The linkage between forecasting and resources provides managers a quantitative basis for hospital capacity planning and management.

As mentioned earlier, medical hours is one of the most scarce and expensive resources in Chilean public hospitals. To estimate the number of medical hours required to attend the expected demand in the emergency service, the following distributions are considered:

- Demand categorization per month, based on illness severity.
- Demand arrival per hour.

For the purpose of this paper, the capacity management analysis, results, and recommendations presented in this section will focus on the emergency service of the Luis Calvo Mackenna Hospital.

4.1. Demand Categorization

Following health care conventions used in Chilean hospitals, patients are classified into 4 categories according to the severity of their illness, as shown in Table 3. Each category is associated to different uses of the medical resource, as will be explained below.

Category	Description
C1	Dying Patient
C2	High Risk Patient
C3	Low Risk Patient
C4	No risk Patient

Table 3: Category Types and Description

As shown in Figure 15, the illness severity distribution varies across the different months of the year. Nevertheless, the severity distribution per month remains relatively stable over the years. Therefore, in the following calculations each month will be considered to have a deterministic distribution of patients for each category.

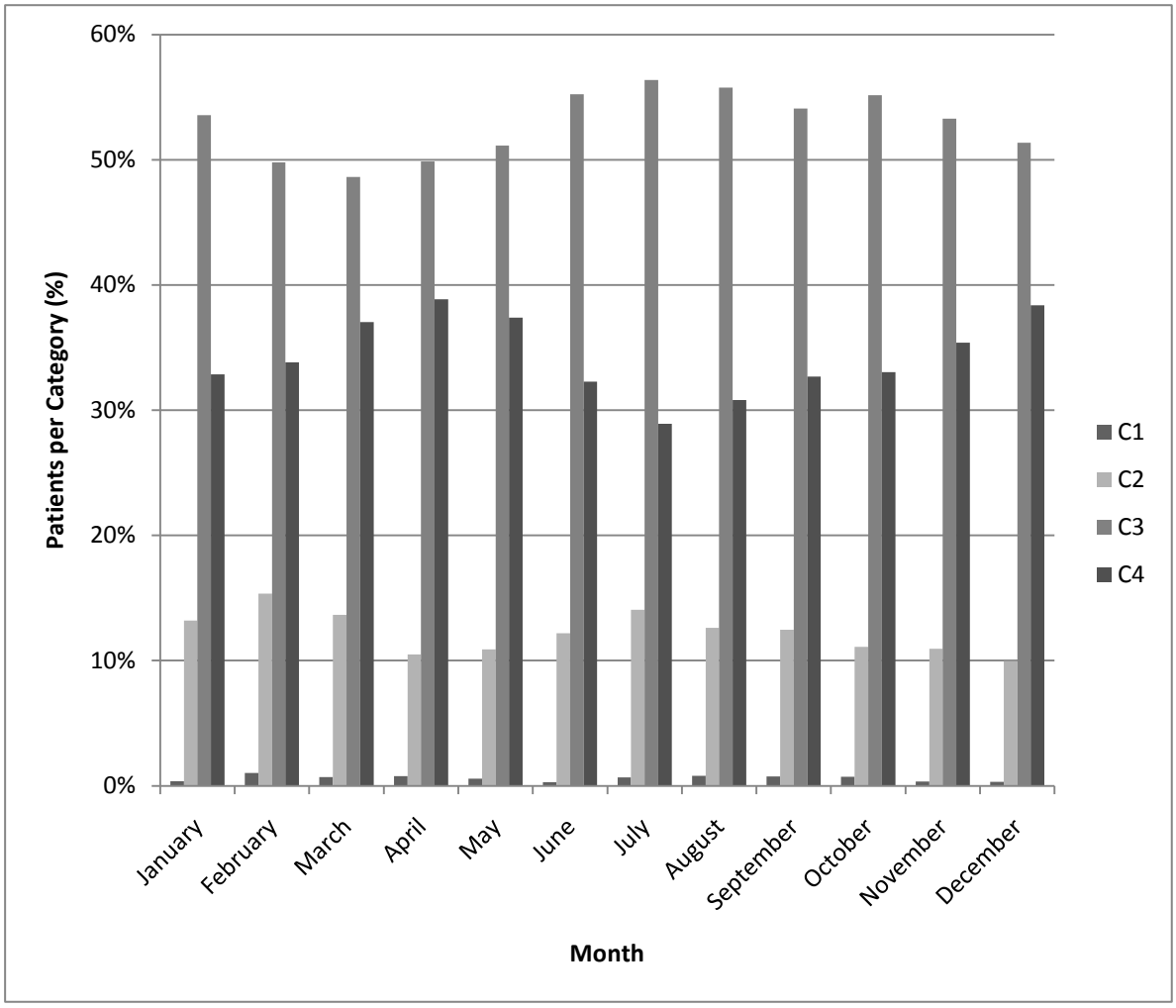

Figure 15: Monthly Categorization Distribution

Given the emergency patients forecast and the illness severity distribution, the expected number of patients per category can be calculated. In order to determine the number of doctors required to attend such demand, the next step is to characterize the behavior of the attention time for each category. For this purpose, a representative sample of C1, C2, C3 and C4 individuals was used.

Each C1 patient is referred to the reanimation room for resuscitation, upon its arrival to the emergency service. When this occurs, and depending on the complexity of the surgery or diagnosis, between one and three the doctors currently working in the attention cubicles leave immediately their activities to focus on the dying patient. After the medical attention, the time required to stabilize and treat the C1 patient is registered in a logbook, along with the names of the doctors that performed the medical procedure. Using this information, we run a Kolmogorov-Smirnov test to determine the distribution of the C1 patients´ attention time. We concluded that it follows a log normal distribution with a mean of 108 minutes and a standard deviation of 121 minutes. We also noticed that the number of doctors required to attend these patients has a

distribution highly concentrated in two doctors, therefore using this value for the following calculations.

When trying to characterize the consults of C2 patients, the data used did not provide enough information to determine a distribution of the attention time. However, in a discussion with doctors a consensus was reached in that the average time to attend C2 patients is 60 minutes, with a standard deviation of 20 minutes. A normal distribution was chosen to represent the behavior of this attention time. Finally, and with a high level of confidence, we determined the distribution of the attention time for C3 and C4 patients as lognormal with means of 10 and 7 minutes, and standard deviations of 7 and 3 minutes, respectively. The non-dying patients are attended by only one doctor.

The time distributions presented above provide a stochastic basis to estimate the time that doctors will spend to attend the patients from each category, expected to arrive to the emergency room. A summary of the attention time distributions found for the different severity categories is presented in Table 4. The number of doctors required to attend each patient per category is presented in Table 5.

Category	Distribution	Mean (min)	Standard deviation (min)
C1	Log Normal	108	121
C2	Normal	60	20
C3	Log Normal	10	7
C4	Log Normal	7	3

Table 4: Distribution of Attention Time per Category

Category	Doctors Req.
C1	2
C2	1
C3	1
C4	1

Table 5: Number of Doctors per Category

4.2. Demand Arrival Distribution per Hour

Patients' arrivals were analyzed at different times of the day, as shown in Figure 16. This study concluded that 59% of the patients arrived at the emergency service from 12:00 until 20:00. Using a representative sample we found that this distribution does not vary significantly among different days of the week nor among the same days of different weeks. Therefore, this distribution is considered as fixed for every day of the year.

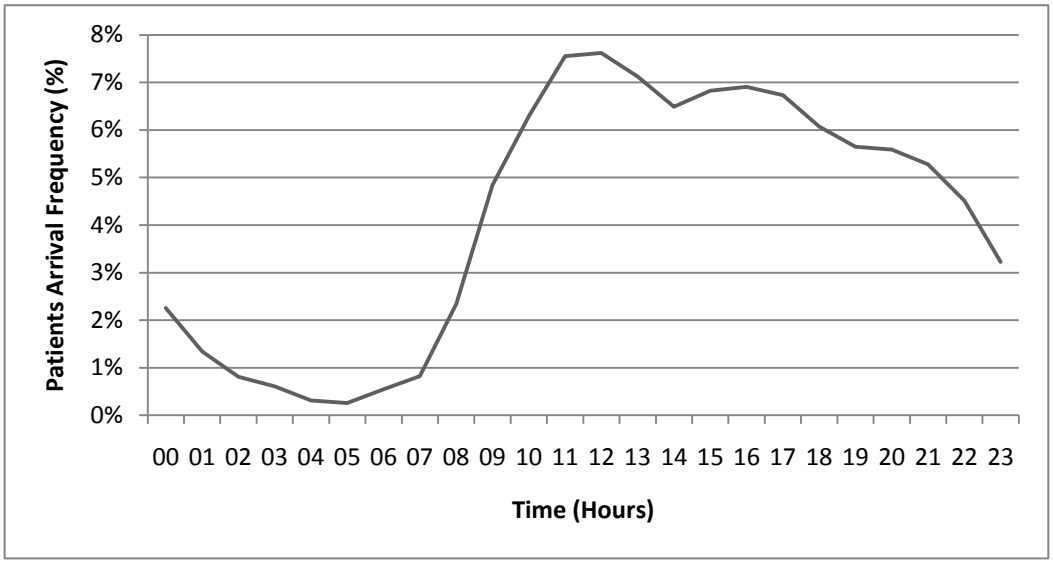

Figure 16: Patients Arrival Distribution per Hour

4.3. Balance of Resources

Once the demand and its behavior have been characterized, the next step was to determine whether the existing resources would suffice to meet the forecasted demand. The unit selected to perform this comparison is medical hours per month.

On the supply side, the total available medical hours per month are determined by the simple multiplication of the number of doctors available by the number and length of shifts per doctor in that period of time. In the case of HLCM, the number of doctors does not vary inter- or intra- shifts. On the demand side, a simple method to calculate the medical hours required to meet the forecasted demand is by considering a deterministic monthly behavior obtained from the model, divided uniformly within each month and distributed within each day as presented in Figure 19. As explained above, the severity distribution of these patients can be considered as deterministic for each month. With these considerations regarding the expected demand, the forecasted number of patients per category can be calculated multiplying the total number of forecasted patients by the proportion of patients per category.

The final step is to convert the demand per category into medical hours, distributed along each hour of the day. To get a quick idea of the medical hours required to meet such demand, we multiplied the forecasted number of patients per category by the mean of the corresponding attention time distributions presented in Table 4. Table 6 illustrates the expected behavior of demand during the day and the availability and rate of use of the medical resources.

Arrival Time	Available Resources [Medical Hours/ Month]	Required Resources [Medical Hours/ Month]	Medical Hours Available [Avail. Resources – Req. Resources]	Occupation Rate (%)
0:00 - 0:59	90	41	49	46%
1:00 - 1:59	90	24	66	27%
2:00 - 2:59	90	15	75	17%

3:00 - 3:59	90	11	79	12%
4:00 - 4:59	90	6	84	7%
5:00 - 5:59	90	5	85	6%
6:00 - 6:59	90	10	80	11%
7:00 - 7:59	90	15	75	17%
8:00 - 8:59	90	42	48	47%
9:00 - 9:59	90	87	3	97%
10:00 - 10:59	90	114	-24	127%
11:00 - 11:59	90	136	-46	151%
12:00 - 12:59	90	138	-48	153%
13:00 - 13:59	90	129	-39	143%
14:00 - 14:59	90	117	-27	130%
15:00 - 15:59	90	123	-33	137%
16:00 - 16:59	90	125	-35	139%
17:00 - 17:59	90	122	-32	136%
18:00 - 18:59	90	110	-20	122%
19:00 - 19:59	90	102	-12	113%
20:00 - 20:59	90	101	-11	112%
21:00 - 21:59	90	95	-5	106%
22:00 - 22:59	90	81	9	90%
23:00 - 23:59	90	58	32	64%

Table 6: Forecasted Medical Resources Occupation Rate

With these simple calculations, several interesting observations arise regarding the medical resources in the emergency service. For example, the period from 0:00 until 8:00 shows a high rate of idle resources, while between 10:00 and 21:00 the medical resources are overly utilized. From the management point of view, for instance, some doctors from the night shift could be reassigned to work during peak hours, but always remaining prepared to meet a potential emergency by having one doctor on duty at home during the night.

4.4. Simulation for Capacity Management in Hospital Facilities

With the calculations from Section 4.3, we were able to determine roughly the performance of the current distribution of resources when attending to the forecasted demand. But the previously mentioned analysis is static and deterministic. In order to provide a more dynamic and accurate perspective of the emergency service performance throughout the attention provided to the patients, we developed a simulation model that incorporates the stochastic behavior of the demand. Since the waiting time and length of lines have shown to be significantly higher for medical attention, the simulation will be performed for these patients only.

The forecast detailed in Section 3 has an error with a normal distribution. To simulate the different demand scenarios for each month, the forecast was adjusted several times by different

values sampled from the normal distribution of the error. Due to the stability of its daily behavior, the demand of each scenario was distributed uniformly across every day of the month. The daily demand was further disaggregated into hourly demand using the distribution shown in Figure 16. As a consequence, we were able to generate several scenarios of monthly demand disaggregated per hour. Using the hourly forecasted demand from each of the scenarios generated as described above, the average forecasted demand was calculated for each hour of the day. We assumed that the hourly demand arrives according to a Poisson process; then this average corresponds to the mean of the Poisson distribution per hour.

Upon their arrival at the emergency service, patients are categorized and served according to the time distributions presented in Section 4.1, which are also stochastic. Now that the stochastic behavior of the demand and the medical attention has been incorporated into the problem, we will discuss the construction of the simulation model and its role in the management of hospital capacity.

In capacity management two types of problems are considered: configuration management, for which we want to determine how different designs of the hospital facilities may affect the quality of service, measured in length of wait before the first medical attention (LOW); and resource management, which decides how current available resources should be assigned to increase the service level, and which and where new resources are required to further improve the quality of service.

The simulation model allows us to observe how the expected flow of patients will use the different services offered in the facilities of the hospital, and how the available capacity performs when attending such demand. As a consequence, capacity can be redistributed or adjusted with the objective of eliminating bottlenecks and reducing idle resources. This provides a powerful decision tool for managing capacity in such a way that a given service level can be guaranteed at minimum cost.

a) Configuration Management

In this section, two main designs of the system will be contrasted with the expected demand: the current configuration and a Fast Track configuration with Triage.

In the current configuration, only dying patients are given priority when arriving at the emergency service. They are immediately taken to the resuscitation service for stabilization, and then referred to the operating room or the intensive care unit service. Patients who are not critically ill must provide their personal data upon their arrival and subsequently wait for medical attention. Therefore, the relative importance in terms of severity of illness is not considered for these patients. The admission time is distributed uniformly between 5 and 10 minutes. After medical attention, the patients can be referred for hospitalization, for diagnostic testing, or immediate discharge. The hospitalization service does not belong to the emergency service and hence does not use its medical resources. The time in diagnostic test services is distributed uniformly between 2 and 4 hours; patients are usually requested to bring the test results back to the doctor within the same day of the evaluation.

The Fast Track configuration includes a Triage to categorize the patients upon their arrival at the emergency service, which is performed by a nurse. The reason behind the creation of the Triage lies in the importance of providing medical attention promptly to patients with the most urgent needs (C2 and C3). Half of the patients categorized as C4 in the Triage are referred to the Fast Track attention, which is performed by a doctor who does educational activities with

medical students from 10:00 to 15:00. This allows reducing the average waiting time within these hours with resources that otherwise would not be used. The rest of the configuration remains as in the current system.

In both configurations outlined above, the doctors shift structure consists of two 12-hour shifts (day and night) with 3 doctors attending in each of them. For each configuration, we developed a simulation model that includes the demand and emergency service characterizations presented earlier. An example of the models is shown in Figure 17.

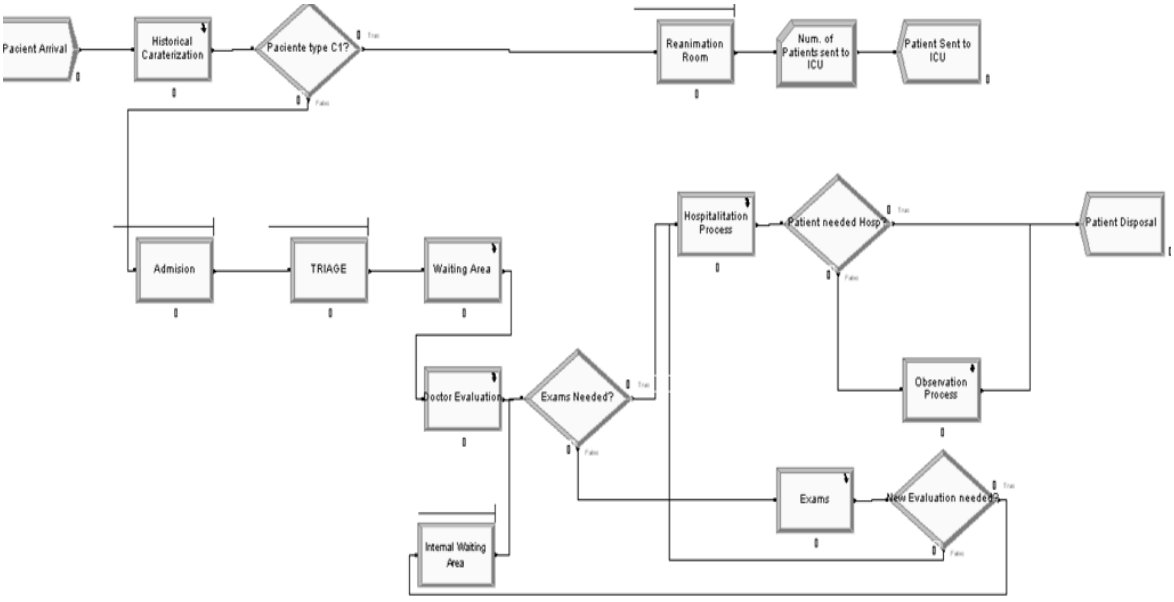

Figure 17: Fast Track with Triage Simulation Model

As stated above, the average LOW will be used as the main criterion to compare the performance of the system designs. This metric was calculated weighing the demand per category by its respective average LOW. The results for this metric for the Base and Fast Track simulated configurations are presented in Table 7.

Configuration (3 doctors day and night)	Avg. (min)	Std. Dev. (min)
Base Case	64.2	1.2
Fast-Track with Triage	57.3	0.9

Table 7: Simulated LOW of Different Emergency Service Configurations

Based on the scenarios run in the simulation, a 95% confidence interval was generated for the LOW of each configuration. The intervals obtained were (55.5, 59.1) and (61.8, 66.6) for the Base Case and Fast Track with Triage, respectively. To test if the LOW differs significantly between these two configurations we applied the procedure proposed by Law and Kelton (1982). This comparison is established based on the difference of their respective statistical distributions, as displayed in Table 8. Since the confidence interval does not contain the value 0, we confirm that the difference shown in Table 7 is statistically significant.

Comparing Configurations	Avg. (min)	Std. Dev. (min)	Lower Bound 95% (min)	Upper Bound 95% (min)
Base Case / Fast Track with Triage	6.9	1.5	3.9	9.9

Table 8: Base Case / Fast Track Configurations Comparison

Using the results presented in Table 7 and 8, we can observe that:

- The simulation model resembles the actual behavior of the system, since current average LOW is within the confidence interval of the simulated Base case.
- The main bottleneck occurs in the medical consults and during the day-shift, as addressed in section 4.3.
- The Fast Track Box with Triage reduces the average LOW in 6.9 minutes, which corresponds to a 10.8% reduction of the current average waiting time.

b) Resource Management

After deciding which configuration performs better on the emergency service, the next question to be addressed is related to the impact that a redistribution, reduction, or addition of medical resources would generate on the performance of the system. The resource management analysis, then, will be performed for the Fast Track configuration only.

The first issue we noticed was that the current shift structure, including the number of doctors per shift, was not constructed based on the daily behavior of the demand, presented in Table 5 and previously in this section. The simulation model was then run for several assignments and number of doctors per shift under the Fast Track configuration with Triage and assuming a stochastic demand.

If the current structure of two 12-hours shifts is maintained, an initial scenario would consider only redistributing the six doctors available in a different manner. Given the greater arrival of patients during the day, as observed in Table 6, a possible redistribution could include the reassignment of a doctor from the night to the day shift. As a consequence, four doctors would attend during the day shift and only two at night. The average LOW of this scenario would be 45.1 minutes.

Further resource management considerations may determine the addition or reduction of medical hours for attending the patients. Since these resources are known to be quite expensive, the different scenarios were simulated by changing the existing capacity in 0.5 doctor intervals. The extra half doctor was included through the creation of a new shift of 6 hours, from 12:00 to 18:00, which is precisely the period in which most patients arrive at the service. Thus, the number of 6.5 doctors available means that 4 doctors attend on the day shift (8:00 - 20:00), 1 doctor on the half shift (12:00 – 18:00), and 2 doctors at night (20:00 – 8:00). The average LOW obtained with this configuration is 40.5 minutes.

The simulation was run using from 5 to 7 doctors within 24 hours, and distributed as explained above. The idea behind analyzing the reduction of the current number of doctors is to assess whether the performance of the system is affected in a significant manner when these resources are lacking, either by management decision or by absenteeism. The average LOW for different numbers and assignments of resources are summarized in Figure 18, including the 95% confidence interval for each of the points.

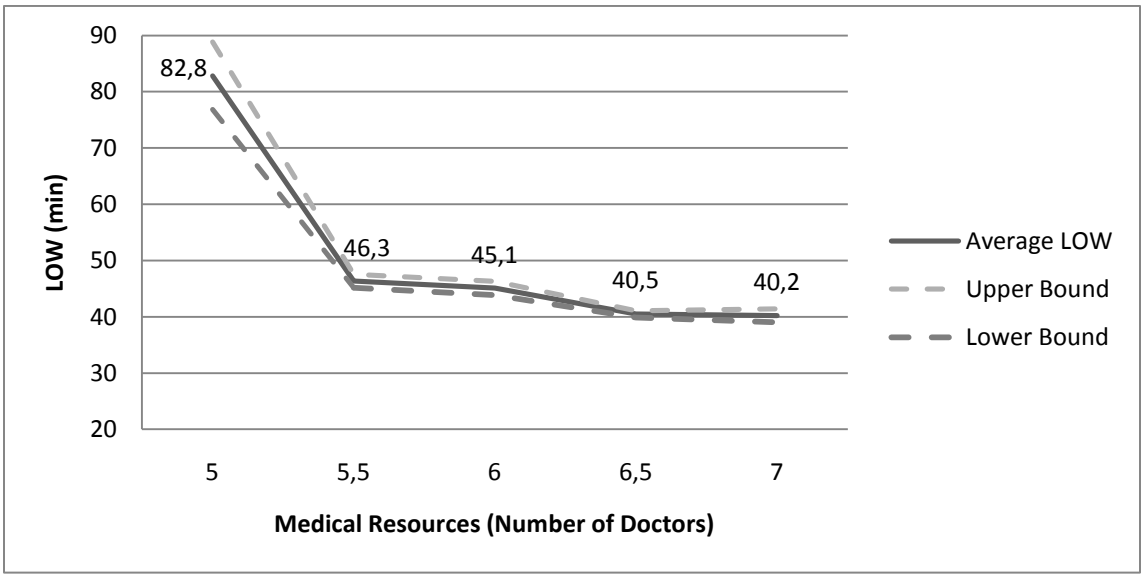

Figure 18: Average LOW and Confidence Intervals for Different Numbers of Doctors

As expected, the addition of medical resources improves the service quality, measured in average LOW. The interesting result is that the average LOW decreases dramatically when reducing the number of medical resources from 5 to 5.5 doctors, while it decreases more gradually when new resources are added. To test whether the LOW difference between all the scenarios included in Table 18 is statistically significant we applied again the procedure proposed by Law and Kelton (1982). Table 9 shows the confidence intervals when comparing the LOW of these scenarios. As it can be observed, increasing from 5 to 5.5 doctors provides a significant improvement to the performance of the system, while the change between 5.5 and 6 doctors does not. Nevertheless, increasing from 5.5 to 6.5 doctors does show statistical significance.

Comparing Scenarios	5.5	6	6.5	7
5	[30.4 ; 42.6]	[31.6 ; 43.9]	[36.3 ; 48.3]	[36.5 ; 48.7]
5.5	-	[-0.4 ; 3]	[4.5 ; 7.2]	[4.4 ; 7.8]
6		-	[3.3 ; 5.9]	[3.1 ; 6.5]
6.5			-	[-1.1 ; 1.6]

Table 9: Confidence Intervals for Compared Scenarios

The previous analysis of the system performance provides hospital managers a decision tool for determining the number and distribution of medical resources on the emergency service, based on a cost/benefit analysis of resources and service improvement.

5. IMPACT OF THE WORK AND CONCLUSIONS

We have shown that the demand for hospital emergency services can be forecasted with great confidence. After testing several forecasting methods we found that SVR provides better results in terms of variance and accuracy. Using this model, we designed a process for managing hospital capacity based on the comparison between the forecasted demand and the available medical resources and a simulation model to assess the performance of different configurations of facilities and resources. Both of these tools assist hospital managers in decision-making regarding the configuration and distribution of medical resources in the emergency service.

The forecasting method and the capacity management process proposed in this paper have been validated and accepted by hospital managers and staff, and are currently in use in one of the hospitals. For this to be possible, we embedded the forecasting model and resources management logic in support computing systems we developed, which may be used in daily work practices. We are currently implementing these processes and systems in one of the hospitals included in this study. The results have been so encouraging that National Health Authorities are considering to extend the demand forecasting and management practices we have designed to close to one hundred public hospitals in Chile.

Acknowledgements: The second author is grateful for the support provided by the Complex Engineering Systems Institute (ICM: P-05-004-F, CONICYT: FBO16) (www.isci.cl).

References

Aburto, L., Weber, R. Improved supply chain management based on hybrid demand forecasts. Applied Soft Computing 2007; 7; 1; 136-144

Adya M, Collopy F. How effective are neural nets at forecasting and prediction? A review and Evaluation. Journal of Forecasting 1998; 17; 451-461

Armstrong J S. (Ed.). Principles of forecasting. Kluwer Academic Publishers: Norwell, MA; 2001

Barros O, Julio C. Application of enterprise and process architecture patterns in hospitals. BPTrends April 2010; www.bptrends.com

Barros O., Julio C., "Enterprise and process architecture patterns", Business Process Management Journal 2011; 17; 4; 598 – 618

Box G E H Jenkins G M, Reinsel G C. Time Series Analysis, Forecasting and Control, 3rd Ed. Prentice-Hall: Englewood Cliffs, NJ; 1994

Chang C C, Lin C J. LIBSVM: A Library for Support Vector Machines[EB/OL], 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Chen P H, Lin C. J., and Schölkopf B., A tutorial on v-support vector machines, Appl. Stoch. Models. Bus. Ind. 2005, 21, 111-136.

Farmer, R D T, Emami, J. Models for forecasting hospital bed requirements in the acute sector. Journal of Epidemiology and Community Health 1990; 44; 307-312

García M L, Centeno M A, Rivera C, and DeCario N, Reducing Time in an Emergency Room Via a Fast-Track, Winter Simulation Conference, 1995, 1048-1053

Hofmann, T, Schölkopf, B, Smola, A. J. Kernel Methods in Machine Learning, Annals of Statistics 2008; 36;1171-1220.

Hwang J, Gao L, Jang W. Joint Demand and Capacity Management in a Restaurant System. European Journal of Operational Research, in press

Jones A J, Joy M P, Pearson J. Forecasting demand of emergency care. Health Care Management Science 2002; 5; 297-305

Kapuscinski R, Zhang R Q, Carbonneau P, Moore R, Reeves B. Inventory decisions in Dell's supply chain. Interfaces 2004; 34; 191–205

Kennedy M, Macbean C E, Sundararajan V, Taylor D M. Review Article: Leaving the Emergency Department without being seen. Emergency Medicine Australasia 2008; 20;4; 306-313

Khurma N, Bacioiu G M. Simulation-Based Verification of Lean Improvement for Emergency Room Process, Winter Simulation Conference 2008, 1490-1499

Law A.M., Kelton W.D. Simulation modeling and analysis. Mc Graw Hill, 2001

Marmor Y N, Wasserkrug S, Zletyn S, Mesika Y, Greenshpan, Carmeli B, Shtub A, Mandelbaum A. Toward Simulation-Based Real-Time Decision-Suppor Systems for Emergency Departments, Winter Simulation Conference, 2009, 2042-2053.

McLaughlin D, Hays J M. Healthcare operations management. AUPHA Press: Washington, DC; 2008; 378-381

MIDEPLAN. Encuesta Casen 2006. www.mideplan.cl/casen

Min D, Yih Y. Scheduling Elective Surgery under Uncertainty and Downstream Capacity Constraints. European Journal of Operational Research, to appear,

Rojas L M, Garavito L A. Analysing the Diana Turbay CAMI emergency and hospitalization processes using an Arena 10.0 simulation model for optimal human resource distribution, Revista Ingeniería e Investigación 2008; 28;1; 146-153.

Samaha S, Armel W S, Stark D W. The Use of Simulartion to Reduce the Length of Stay in an Emergency Department, Winter Simulation Conference, 2003, 1907-1911

Schweigler L M, Desmond J S, McCarthy M L, Bukowski K J, Ionides E L, Younger J G. Forecasting models of emergency department crowding. Academic Emergency Medicine 2009; 16; 301-308

Shirxia Y, Xiang L, Li N, Shang-dong Y. Optimizing neural network forecast by immune algorithm. School of Business Administration, North China Electric Power University, Beijing, China; 2007; http://www.springerlink.com/content/t23961t0x38nvu3h/

Smola A J, Schölkopf B. A tutorial on support vector regression. Statistics and Computing 2004; 14; 3

Vapnik V. The Nature of Statistical Learning Theory, Springer-Verlag, 1995

Vapnik V. Statistical Learning Theory, Wiley, 1998

Vladimir C, Ma Y Q. Practical selection of SVM parameters and noise estimation of SVM regression. Neural Networks 2004; 17;1; 113 126

Zhang G P. Avoiding pitfalls in neural network research. IEEE Transactions on Systems, Man and Cybernetics—Part C: Applications and Reviews 2007; 37; 3-13