ESANN'2005 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 27-29 April 2005, d-side publi., ISBN 2-930307-05-6.

Predicting Bed Demand in a Hospital using Neural Networks and ARIMA models: a Hybrid Approach

Mark P. Joy and Simon Jones
School of Mathematics, Kingston University,

Kingston-upon-Thames, Surrey KT1 2EE, UK.
m.joy@kingston.ac.uk

Abstract. In this paper we describe an investigation into the prediction of emergency bed demand - bed demand due to non-scheduled admissions - within a NHS[1] hospital in South London, U.K. A hybrid methodology, incorporating a neural network and an ARIMA model was used to predict a time series of bed demand. A thorough statistical analysis of the data set was performed as a preliminary phase of the research from which a classical linear predicting model was developed. The prediction errors or residuals from this model were then used as input to a neural network. These methods represent a novel approach to the problem of efficient bed resource management for hospitals.

1 Introduction

NHS hospital bed mangers face a difficult task in attempting to allocate their beds between emergency admissions and so-called elective admissions – those admissions that are planned and, in general, referred by the patient's doctors or consultants. Depleting the bed stock 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. Forecasts of demand are desperately needed for bed management in the NHS. The report [1] identifies that around a quarter of Trusts[2] make no assessment of likely bed demand for more than a day ahead. Moreover, whilst the National Audit Office acknowledges that it is difficult to predict emergency admissions with precision, it strongly urges NHS Trusts to *"make more effective use of their knowledge of patterns of emergency admission to assess the likely demand on their resources"*, [1]. In this paper we describe an investigation into the prediction of emergency bed demand carried out at an acute hospital in South London, U.K.

Time series data consisting of approximately 5 years of daily maximums, aggregated on a weekly basis, of emergency bed occupancy was released to us by the research sponsor – Bromley NHS Trust. A central goal of the research is

[1]In the UK the state provider of health is the National Health Service – abbreviated to NHS throughout this paper.

[2]Within the NHS a Trust is a collection of hospitals with financial autonomy, charged with implementing governmental health policy and accountable directly to government.

to develop a reliable means of predicting peak or maximum emergency demand on a weekly basis so that more informed and reliable strategic decisions can be made by bed managers without committing unacceptably high risks of depleting the bed stock for emergency cases.

We report here only on the one-step problem for our data. This means that we consider the problem of developing a model for the prediction of the weekly maximum for one week hence. Discussions with bed managers reveal that this is likely to be an extremely useful indicator and theoretically will allow planners to allocate bed resources in a more efficient manner by anticipating the likely demand on bed stock during the following week.

The research reported here builds upon that presented in [2], which can be considered as the starting point of this investigation.

2 Methodology

There are perhaps many competing 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. Central to such a study is Takens embedding theorem [3] which loosely says that there is an integer d such that the dynamics of the time series can be reconstructed by embedding the series as a sequence of vectors $(x_1, \ldots, x_d) \in \mathbf{R}^d$. Our chosen methodology for estimating d is presented in the next section.

In summary we investigated many models (in particular regime switching models of the SETAR variety and an FIR neural network with temporal back-propagation) but we settled upon a hybrid one, prompted by [4], wherein a forecasting model consists of a stochastic linear model, $L_t = L(x_t)$, and a neural network, N, such that if e_t is the residual of the fitted linear model L,

$$e_t = x_t - L_t$$

then the hybrid forecast will be

$$x_t = L_t + N_t(e_t).$$

We followed a classical Box-Jenkins approach to the fitting of the linear model (see [5]), very briefly summarised in the next section.

Perhaps the final justification of the chosen model is the accuracy of out-of-sample predictions. In fact our model contains considerable predictive power.

3 Model Development

3.1 Linear Model Development

The material in this section will be very brief, we refer the reader to [5] for an exposition of the theory and practice of fitting linear models of the ARIMA type.

The most fundamental linear, stochastic model of a time series which encapsulates dependencies on historical values (the autoregressive element) as well as dependencies on independent random disturbances is the ARMA model

$$x_t = \sum_{i=1}^{p} \psi_i x_{t-i} + \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}. \tag{1}$$

The $p + q + 2$ parameters $(\mu, \theta_1, \ldots, \theta_p, \psi_1, \ldots, \psi_q, \sigma^2)$, need to be estimated from the data set, (μ is the mean and σ^2 the variance of the independent and identically distributed (i.i.d) random variables ε_t). However, ARMA models assume an underlying stationary time series. In practice some kind of filter is applied to x_t to render the series stationary. Quite often differencing does the trick; we define a new time series $y_t = x_t - x_{t-1}$, and iterate this differencing operation until the series y_t cannot be distinguished from a stationary series by significance testing. An ARMA model of order p, q that has been differenced d times to render it stationary is known in the trade as an ARIMA(p, d, q) model. Further details of model fitting are omitted due to lack of space.

The linear model fitted onto the provided data set is an ARIMA(1,1,2) (strictly speaking it is a SARIMA model, essentially two ARIMA models, capturing the weekly and seasonal statistics of the time series. The seasonal component being (0,1,0), in other words we have differenced once to render the filtered series annually stationary).

3.2 Lagged Mutual Information

An alternative to studying the ACF of a time series is the estimation of the lagged mutual information. The definition of this measure is

$$S(\tau) = -\sum_{ij} p_{ij}(\tau) \log\Big(\frac{p_{ij}(\tau)}{p_i p_j}\Big) \tag{2}$$

where, for some partition on the real numbers, p_i is the probability of finding a time series value in the i-interval, and $p_{ij}(\tau)$ is the joint probability that an observation that falls into the i-th interval falls into the j-th interval τ time units later. In theory this expression has no systematic dependence on the size of the partition elements and can be quite easily computed. There exist good arguments (see [6]) that if S exhibits a marked minimum for a value of τ, then this is a good candidate for a embedding dimension guaranteed by Taken's theorem to reconstruct the dynamics. In practice we use this measure to determine the size of the input layer to the neural network trained on the residuals. For more details and available software for calculating S, see [7].

4 The Neural Network Model

We report here on a candidate neural network with an $8-3-1$ architecture (that is, 8 input units, 3 hidden units and one linear output unit); for a discussion of

alternative architectures see section 5. The number of inputs was obtained as the first minimum of the lagged mutual information function.

Of importance here of course is the algorithm used to train the network. We used the GRG2 algorithm – a modified version of the generalised reduced gradient method (see [8]). The superiority of GRG2 over stochastic gradient descent in quickly finding better solutions over a broad range of optimisation problems has been reported by several researchers, see for example, [9].

We adopted the *Statistical Stepwise Method* (SSM) for weight elimination, [10]. We prefer this method to competing approaches because of it's statistical rigour. If we unpack the estimated network weights into a vector, $\hat{W} = (\hat{w_1}, \ldots, \hat{w_m})$, then the SSM algorithm consists of 4 steps:

1. Compute the measures $Q(l) = \frac{\hat{w_l}}{\hat{\sigma}(\hat{w_l})}$, where $\hat{\sigma}(\hat{w_l})$ is the estimated standard deviation of w_l (computed using the inverse Hessian matrix, see [10], p.1357);

2. Define $l_* = \text{argmin}_{w_l \in W}\{Q(l)\}$, – the argument of the minimum value of the quotients in Step 1;

3. Define W_l as the set of weights W with $w_l = 0$, then test the model W_{l_*} (null hypothesis) against the model $\hat{W}$ (the alternative hypothesis): accept the elimination of w_{l_*} if $Q(l_*) < \alpha$, for some critical value α. This significance test is a Student t-test and we set $\alpha = 1.96$. Actually, as recommended in [10], we use $\alpha = 1.5$, in accordance with most stepwise regression tests in statistics.

4. In case we reject the null hypothesis, stop the process and keep the previous set of weights. Otherwise retrain the network corresponding to the model W_{l_*} and go to Step 1.

The SSM method is theoretically superior to the OBD method defined in [11], which centres around the notion of the *saliency*, s_l, of a weight w_l. The saliency of a weight is the increase in residual error that results from its elimination: $s_l = (1/2T)(S(\hat{W}_l) - S(\hat{W}))$, here $S(W)$ is the sum of squared residuals using the set of weights W. Now it can be shown that

$$Q(l)^2 = \frac{2Ts_l}{\frac{S(W)}{T-m}}$$

and the asymptotic distribution of $Q(l)^2$ (as $T \to \infty$) is a χ^2-distribution with one degree of freedom. Thus although significance testing the $Q(l)$ statistics is equivalent to the comparison of the saliency statistics, the drawback with the OBD scheme is that it does not take into account the relative value of the saliency. Moreover the use of only diagonal elements of the Hessian matrix in OBD is equivalent to assuming that weights are independentely distributed. The SSM algorithm, makes no such assumptions and uses the statistic $Q(l)$ directly, thereby capturing the notion of relative size of each saliency. For more details, see [10].

A training set of 500 observations of weekly maximum occupancy was divided into in-sample training sets and out-of-sample training sets. The mean occupancy for the observed period was 411.8 beds (median 411), with a standard deviation of 27.8. The values fluctuated exhibiting a minimum of 335 and a maximum of 526. The inter-weekly fluctuation, measured by the standard deviation weekly differences (i.e. the first difference of the time series) was 14.5, approximately 3.5% of total occupancy; the data was normalised before presentation to the hybrid model.

We measure the performance of both models by the out-of-sample RMSE, defined as

$$\text{RMSE}^2 = \frac{1}{T}\sum_{t=1}^{T}(\hat{x}_{t+1} - x_{t+1})^2,$$

where $\hat{x_t}$ is the model prediction at time t and T is the size of the out-of-sample training set.

We summarise our findings in the table below. Table 1 provides the in-sample and out-of sample performance of the RMSE for the ARIMA and hybrid models.

	num. of (8)-vectors	ARIMA	Hybrid	imp.
in-sample	350	22.439	16.995	24.3%
out-of-sample	140	23.009	19.96	13%

Table 1: RMSE comparisons.

The last column measures the improvement in the hybrid predictions compared to the ARIMA model. We see that in-sample predictions are superior and out-of-sample improvements are much reduced but still provides 13% improvement. There is little degradation in performance in prediction for the linear model from in-sample to out-of-sample performance, however the hybrid model has larger out-of-sample RMSE. As for weight elimination the SSM algorithm eliminated, for example, 3 input-to-hidden weights from the fully connected $8-3-1$ model and improved out-of-sample forecasting by approximately 4%. The findings of this research suggest that weight elimination according to SSM is largely insensitive to the size of the critical value α constrained to lie in the interval $(1, 1.96)$ say, since the weight eliminations that took place were all confirmed by values of $Q(l)$ much smaller than 1.0.

5 Conclusions

We have presented a time series prediction problem concerning bed occupancy in a hospital in the UK. In the context of an increasing awareness for cost-efficiency in the provision of hospital services in the UK this is an important application; it is, to these authors' knowledge, also a novel application. We have been able to demonstrate the improvement which results from employing a hybrid model consisting of a linear ARIMA predictor and a neural network, trained to capture the residual nonlinear correlations undetected by the linear

model. Improvements are substantial as evidenced by both in-sample and out-of-sample RMSE. The model described in this paper is currently being bench-tested in the hospital in order to assess the accuracy and usefulness of such a tool for everyday use by bed managers. Alternative neural network architectures are also under consideration but the model reported here is currently the best, measured by out-of-sample RMSE.

One future direction for this research is to consider the possiblity of the presence of regimes in the time series data. Wavelet analysis of daily occupancy in the hospital reveals the existence of a 7 day cycle that exhibits intermittency. Thus there is evidence that *regime-switching* occurs within this system. Indeed, one regime appears to coincide with increased demands on bed supply during the winter months. However, models constructed according to this paradigm performed poorly in out-of-sample tests compared to the ARIMA model and hence to the hybrid model. We plan to consider the possibility of building hybrid models h_R for the distinct regimes, $R = 1, 2$, and switching between them at the intermittency times of the 7-day cycle. It is conjectured that such a scheme will provide more finely tuned and therefore better forecasts.

References

[1] *Inpatient admissions and bed management in NHS acute hospitals*, National Audit Office, 2000.

[2] Simon Andrew Jones and Mark Patrick Joy, *Forecasting demand of emergency care*, Health Care Mangement Sci., 5, 2002, pp. 297–305.

[3] Takens F., *Detecting the strange attractors in turbulence*, in D. A. Rand and L. S. Young (ed.s), Dynamical Systems and Turbulence, Lecture Notes in Mathematics, 898, New York: Springer-Verlag, 1980.

[4] G. Peter Zhang, *Time series forecasting using a hybrid ARIMA model and a neural network*, Neurocomputing, vol. 50, Jan. 2003, pp. 159–175.

[5] G. E. P. Box and G. Jenkins, Time Series Forecasting and Control, Holden-Day, San Franciso, CA, 1970.

[6] A. M. Fraser and H. L. Swinney, *Independent coordinates for strange attractors from mutual information*, Phy. Rev. A 33, 1986.

[7] R. Hegger, H. Kantz and T. Schreiber, *Practical implementation of nonlinear times series methods: The TISEAN package*, Chaos, 9, p. 413, 1999.

[8] Abadie, J., and Carpentier, J., *"Generalization of the Wolfe reduced gradient method to the case of nonlinear constraints"*, in: R.Fletcher (ed.), Optimization, Academic Press, London, 1969, pp. 37–47.

[9] Ming S. Hung and James W. Denton, *Training neural networks with the GRG2 nonlinear optimizer*, European J. of Oper. Res., 69, 1993, pp. 83–91.

[10] Marie Cottrell, Bernard Girard, Yvonne Girard, Morgan Mangeas and Corinne Muller, *Neural modeling for time series: a statistical stepwise method for weight elimination*, IEEE Trans. on Neural Networks, vol. 6, No. 6, pp. 1355–1363, Nov. 1995.

[11] Y. Le Cun, J. S. Denker and S. A. Solla, *"Optimal Brain Damage"*, in D. S. Touretzsky, (ed.), Advances Neural Information Proc. Syst. II, San Mateo, CA: Morgan Kaufman, 1990, pp. 598–605.