Available online at www.sciencedirect.com
ScienceDirect
European Journal of Operational Research 180 (2007) 354–368
EUROPEAN JOURNAL OF OPERATIONAL RESEARCH
www.elsevier.com/locate/ejor

Stochastics and Statistics

Forecasting with cue information: A comparison of multiple regression with alternative forecasting approaches

K. Nikolopoulos [a], P. Goodwin [b,*], A. Patelis [c], V. Assimakopoulos [d]

[a] *Decision Sciences and Operations Management Group, Optima Division, Manchester Business School, University of Manchester, M15 6PB Manchester, United Kingdom*
[b] *The Management School, University of Bath, Bath BA2 7AY, United Kingdom*
[c] *Forecasting Systems Unit, School of Electrical and Computer Engineering, National Technical University of Athens, 9, Iroon Polytechniou Street, 15773 Zografou, Athens, Greece*
[d] *Secretary for the Information Society, Ministry of Economy and Finance, 5–7 Nikis Street, 10180 Athens, Greece*

Received 16 November 2004; accepted 17 March 2006
Available online 9 June 2006

Abstract

Multiple linear regression (MLR) is a popular method for producing forecasts when data on relevant independent variables (or cues) is available. The accuracy of the technique in forecasting the impact on Greek TV audience shares of programmes showing sport events is compared with forecasts produced by: (1) a simple bivariate regression model, (2) three different types of artificial neural network, (3) three forms of nearest neighbour analysis and (4) human judgment. MLR was found to perform relatively poorly. The application of Theil's bias decomposition and a Brunswik lens decomposition suggested that this was because of its inability to handle complex non-linearities in the relationship between the dependent variable and the cues and its tendency to overfit the in-sample data. Much higher accuracy was obtained from forecasts based on a simple bivariate regression model, a simple nearest neighbour procedure and from two of the types of artificial neural network.

Keywords: Forecasting; Regression; Neural networks; Judgment; Analogies

1. Introduction

Multiple linear regression (MLR) is a common choice of method when forecasts are required and where data on several relevant independent variables (or cues) is available. The technique has been used to

* Corresponding author. Tel.: +44 122 5323594; fax: +44 122 5826473.
E-mail addresses: kostas.nikolopoulos@mbs.ac.uk (K. Nikolopoulos), mnspg@bath.ac.uk (P. Goodwin).

0377-2217/$ - see front matter
doi:10.1016/j.ejor.2006.03.047

produce forecasts in a wide range of areas (e.g. Burger et al., 2001; Sadownik and Barbosa, 1999; Chu and Zhang, 2003) and there is evidence that it is often used by companies to derive forecasts of demand from marketing variables and various macroeconomic measures (e.g. Mentzer and Bienstock, 1998). It is a method that is covered in almost all textbooks on business statistics and forecasting (e.g. Makridakis et al., 1998) and it is a standard fixture of MBA and undergraduate courses in data analysis and forecasting. However, recent evidence by Gigerenzer and Todd (2000) has shown that judgmental forecasters employing very simple heuristics can often outperform multiple linear regression when forecasts of binary variables are required. Similarly, earlier work by Dawes and Corrigan (1974) demonstrated the robust accuracy of simply applying equal weights to the values of the independent variables in order to obtain forecasts. This occurs because the likelihood function is often fairly flat over many different combinations of weights (Astebro and Elhedhli, 2003).

There are a number of possible reasons why MLR might be expected to lead to less accurate forecasts than alternative methods. First, MLR models may be too complex in that they include representations of false patterns in the noise associated with past data. For example, Gigerenzer and Todd (2000) have suggested that MLR tends to lead to models which are overfitted to past observations so that hold-out sample forecasts are relatively inaccurate. Indeed, this must account for situations where heuristics using single independent variables outperform MLR since the single-cue heuristic is a nested model of a multiple regression model (Astebro and Elhedhli, 2003). The theoretical circumstances where simple heuristics may be expected to outperform MLR in a binary choice problem have been identified by Hogarth and Karelaia (2004).

At the other extreme, MLR models may be too simple. Because they are, by definition, linear they will be unable to represent non-linear relationships between variables. For example, this may occur in situations where the outcome depends on a non-compensatory combination of cue values (e.g. a low score on one cue might determine the outcome irrespective of whether the other cues have high or low values) or where outcomes depend upon products of some of the cue values rather than their weighted sums. A third possibility is that changes in the environment mean that the structure of the data used to derive the MLR model differs from that that which applies in the forecasting periods. In these periods, new relationships between the outcomes and cues might apply or special, rare, events may occur which mean that the normal relationships are temporarily suspended (these are referred to as ''broken leg'' cues).

These possible causes of the relatively poor performance of MLR models suggest that it is may be worth considering alternative methods. If the models are too complex then the most obvious alternative would be to use a simple bivariate regression model based on the cue that is most highly correlated with the dependent variable. Another alternative could involve simply selecting a set of past cases that have the most similar values for the independent variables to those that will apply in the forecast period. The mean value of the dependent variables for these cases can then be used as the forecast. In this paper this method will be referred to as ''nearest neighbour analysis (NN)''. Although the method is conceptually simple, the computation required in order to identify the most similar cases to the one in question can be quite challenging. An obvious alternative to NN involves directly using human judgment, rather than attempting to replicate it. Human judges can potentially spot ''broken leg cues'' and environmental change and they may also be able to handle complex configural information (information imbedded in the pattern of the cues), though some evidence casts doubt on this (Shanteau, 1992). Unlike statistical methods, they can also potentially make use of qualitative cues, such as verbal descriptions of forthcoming events. However, judges also have limited information processing capacity, which often causes them to over simplify problems, to produce biased judgments and to behave inconsistently (Goodwin, 2002).

On the other hand, if MLR models are too simple to represent complex relationships then forecasts based on artificial neural networks (ANN) may lead to improved accuracy. ANNs can learn about and, hence model, complex non-linear relationships. However, overfitting is a potential consequence of this complexity (Chatfield, 1993).

Note than an alternative to the use of the single forecasting methods that we have described involves obtaining forecasts by combining two or more of the methods. Combining, which often entails taking a simple average of forecasts derived from different methods, has been shown to be effective in a number of studies since the seminal work of Bates and Granger (1969) (also see Clemen, 1989 for a review). Combining is thought to improve forecast accuracy because the constituent forecasting methods draw on different information to produce their forecasts. Where the methods are based on the same information, or suffer from the same biases, there are therefore likely to few gains in accuracy to be obtained from combining. Indeed, where the combination involves a simple average, it can be shown that the increased accuracy resulting from combination depends upon the correlation of the forecast errors of the constituent methods – the *lower* the correlation then the higher the expected accuracy will be (Goodwin, 2000). Since the objective of the current paper is to examine the reasons for the relative performance of the individual methods, combination will not be discussed further.[1]

This paper evaluates the relative performance of these alternative forecasting methods (including different versions of some of the methods) by applying them to the problem of forecasting the share of audiences achieved by TV networks when broadcasting sporting events. The problem is an important one –for example advertisers will base their plans on such forecasts – and it lends itself to the application of techniques like MLR. Our emphasis in the paper is on the comparison of the methods applied in an automated form such that they require little expert supervision. This approach is useful when a large number of forecasts need to be made on a regular basis. Moreover, in a non-automated form the quality of the resulting models is often dependent on the personal judgments and expertise of the analyst and hence the results may not be easily generalised. The paper is structured as follows. The next section provides a more detailed discussion of the methods, together with a description of their application to TV audience data obtained from Greece. This is followed by an analysis of the relative accuracy of the methods, together with an assessment of the reasons underlying their different levels of performance. The paper concludes by inferring some general principles for choosing forecasting methods when data on several cues is available.

2. The methods and how they were applied to the data

2.1. Data

The dataset used in this study is television (TV) audience ratings from 1996 to 2000 in Greece. TV audience ratings are often measured as a percentage (%) of people watching a specific programme relative to all those having a TV set on at a given time. When the TV programme broadcasts a special event, such as a major sport event or a film premiere, the audience level is radically affected. Usually there is an increase in audiences prior to, and during, the special event.

Impact refers to the additive increase in the percentage of the total audience above the average of the last four relative periods (same time zone within day, same day of week) for a given channel. For example, suppose that a programme broadcasting a special sport event has a rating of 60% on a Wednesday night from 10.15 p.m. to 10.30 p.m., and the average of the last four Wednesdays for that channel in the same time zone was 45%. In this case the additive impact is $60\% - 45\% = 15\%$. It is also possible to define impact in terms of particular groups of people who may, for example, be targeted by advertisers, such as adult males in a particular age group. If this is the case, impact will be defined in relation to that group of people, rather than the total audience.

[1] In the current study the correlations between the post-sample forecast errors of the pairs of different methods all tended to be high and positive – the lowest correlation was 0.724 – anyway. Not surprisingly therefore, we found that combination did not lead to significant gains in accuracy when we investigated it.

Table 1
Sporting events: parameters (Independent variables)

Variable	Name	Range	Description
1	Importance	1(very low) to 5(very high)	Describes the importance of the sport event for the Greek spectators
2	Competition	1(low) to 3(high)	Describes the level of competition provided by the shows on the alternative channels
3	Time Zone	1(low) to 3(high)	Describes the importance of the time zone within the day

Table 2
Sample of the dataset

Id	Date	Start	End	Main title	Secondary title	Impact (%)	Importance	Competition	Time zone
1	15/4/1998	21:39	23:44	Football unclassified	Dortmund–real Madrid	19	3	2	3
2	15/4/1998	24:21	25:59	Football unclassified	Monaco–Juventus	12	3	2	2
6	20/5/1998	21:32	23:50	Champions league	Juventus–real Madrid–final	43	4	2	3
8	28/6/1998	17:16	20:05	World Cup 1998	France–Paraguay	49	3	2	2
16	9/8/1998	19:51	22:13	MUNDOBASKET 98	Russia–Yugoslavia	13	3	3	3
18	21/10/1998	21:38	23:45	Champions league	Olympiakos–Ajax	?	2	1	3

AGB Hellas S.A. provided for this study real TV audience ratings data, relating to the target group of *men* aged 25–44 in Greece. The supplied data included 46 sport programmes (special events). Each event is described by three categorical parameters (independent variables). These parameters, *Importance*, *Competition* and *Time Zone* are described in detail in Table 1. The parameters are quantified on scales from 1 to 3 or 1 to 5 in order to make statistical forecasting models easily applicable.

From the 46 programmes for which data was provided, 34 were used for the fitting or training the models and 12 were used as a hold-out sample to evaluate forecasting performance. The hold-out sample was selected to ensure that it included a range of events (e.g. basketball and football) for which there were similar events in the in-sample data, otherwise the methods would have been required to produce forecasts for some types of events without access to their "history". The values for the three independent variables (cues) were estimated by Dr. A. Patelis, an expert in audience analysis in Greece. A subset of this data is given in Table 2. The first five rows in Table 2 give examples of programmes that were used for fitting and training and the sixth gives an example of a programme that was used in the hold-out sample. There was no significant difference between the mean impact of programmes in the in-sample data and those in the hold-out sample ($p = 0.12$). The impacts for the in sample data ranged from 2% to 64% while those for hold-out sample ranged from 4% to 47%. Nor were there any significant differences in the mean values of the independent variables (minimum p-value = 0.20). Only one of the 12 forecasts for the hold-out sample involved an extrapolation in that Importance for this event had a value of 1, while the range of values for Importance in the in-sample data was 2–5.

2.2. Methods

In this section, the four different approaches that were used to forecast the impact of sport events are discussed in detail.

2.2.1. Nearest neighbours

Nearest neighbour approaches (Härdle, 1992), are a very simple way of estimating the impact of future events by looking at the past for the impact of similar or analogous events (Green, 2002). These approaches

have been applied with mixed success to time series data to obtain forecasts in a wide variety of domains, including foreign exchange rate forecasting (e.g. Meade, 2002), stock market forecasting (Kanas, 2003; Fernandez-Rodrguez et al., 1999), avalanche forecasting (Brabec and Meister, 2001) traffic forecasting (Sun et al., 2003). and market response forecasting (Mulhern and Caprara, 1994).

2.2.1.1. Finding the neighbours. This task requires ranking historic events in terms of *similarity* (these are usually referred to as *fitting* or *learning* events). Similarity requires a metric of *distance*. Various metrics have been used in order to measure distance in multidimensional space, including Euclidian norms (Härdle, 1992). Multidimensional space metrics are required since our events involve three parameters (*Importance*, *Competition* and *Time Zone*). For simplicity's sake, a very straightforward heuristic was introduced, based on the Absolute Percentage Error (APE) (Makridakis et al., 1998). This was used to measure the distance between similar events as shown below:

$$\text{Distance}_{\text{Ev1,Ev2}} = \text{Abs}((I_{\text{Ev1}} - I_{\text{Ev2}})/I_{\text{Ev2}}) + \text{Abs}((C_{\text{Ev1}} - C_{\text{Ev2}})/C_{\text{Ev2}}) + \text{Abs}((T_{\text{Ev1}} - T_{\text{Ev2}})/T_{\text{Ev2}})$$

where Evk = Event k, I = *Importance*, C = *Competition* and T = *Time Zone*.

2.2.1.2. Forecasting impact. Three different methods have been used to forecast the impact of future events:

- *NN*
  This is the simplest approach. The nearest neighbour provides the forecast of the impact for the event under consideration.
- *3-NN*
  In this version the three nearest neighbours are used for the forecast of the impact for the event under consideration. The forecast is a weighted average with the impact of the nearest neighbour weighted 50% and the second and third nearest neighbours each weighted 25%. To be mathematically correct, in this case a triangular kernel function is used (Härdle, 1992), assigning weights equal to (1/2, 1/4, 1/4) to the three nearest neighbours.
- *k-NN*
  In this version five nearest neighbours are used, and their impacts are equally weighted (20%) in order to obtain the forecast of impact for the event under consideration.

2.2.2. Regression

Classical least squares was used to obtain three different linear regression models (Makridakis et al., 1998). To compare regression on an equal basis with the alternative methods, our emphasis was on testing the accuracy of regression methods that are the most commonly used in forecasting and which require the minimum expert intervention in the derivation of the model. In each case the model parameters were estimated from the data on the 34 in-sample TV programmes.

- *R-all*
  In this model all of the independent variables were used in the model, irrespective of the significance of their partial regression coefficients. The model obtained is shown below (the p-values are displayed in brackets below these coefficients for information):

$$\begin{aligned}\text{Impact} &= 85.3 + 0.54 * \text{Importance} - 11.6 * \text{Competition} - 10.5 * \text{Time Zone}\\ &\quad (0.00)\ (0.85) \qquad\qquad (0.00) \qquad\qquad (0.01)\end{aligned}$$

$$R - \text{squared} = 41.9\%.\ \text{Adjusted } R - \text{squared} = 36.1\%$$

- *R-step*
  To obtain this model standard stepwise regression was employed to identify and estimate the model. The resulting model is shown below:

$$\text{Impact} = 83.7 - 11.6 * \text{Competition} - 10.6 * \text{Time Zone}$$
$$(0.00)\ (0.00) \qquad\qquad (0.01)$$
$$R - \text{squared} = 41.9\%.\ \text{Adjusted}\ R - \text{squared} = 38.1\%$$

It can be seen that the removal of *Importance* from the model has had little effect on the coefficients of the other variables

- *R-pars*
  This is the most parsimonious model and is based only on the independent variable which had the highest correlation with impact. The model is:

$$\text{Impact} = 55.8. - 12.0 * \text{Competition}$$
$$(0.00)\ \ (0.00)$$
$$R - \text{squared} = 26.2\%.\ \text{Adjusted}\ R - \text{squared} = 23.8\%$$

This type of model might be attractive when the cost of collecting data on several independent variables is high. However, it can be seen that the model has the poorest fit to the in-sample data.

Diagnostic checks on all the models suggested that there was no heteroscedasticity and that the residuals were approximately normally distributed. For the two multiple regression models there was no evidence of multicollinearity.

2.2.3. Artificial neural networks

A fully connected multilayer perceptron with two hidden layers and an output layer was used for the impact estimation (Haykin, 1998). Similar networks have been utilized in order to forecast the impact of future events (Nikolopoulos and Assimakopoulos, 2003; Lee and Yum, 1998). In our case the size of the input signal was equal to three (that is the number of parameters explaining the historic sport events). Each input represented one parameter/independent variable. The size of the output signal was equal to 1. In order to forecast the impact of a possible future event, the network was trained first with the training set (34 events), using the Back Propagation algorithm (BP, Haykin, 1998). Each sport event was considered to be one training example and its impact was considered to be the desired response of the network. Then 12 hold-out events were fed into the network with the form of an input signal. Signals flowed through the network in a forward direction, from left to right and on a layer-by-layer basis reaching the output node. The outcome was the predicted impact of the event.

2.2.3.1. Limitations of the BP algorithm. Since back-propagation is basically a hill climbing technique, it runs the risk of being trapped in a *local minimum* where every small change in synaptic weights increases the cost function. It is clearly undesirable to have the learning process terminate at a local minimum, especially if it is located far above a global minimum (Haykin, 1998). One possible solution could involve the user retraining the network until it converges to a desired local minimum. Of course the global minimum of the error surface is not known a priori so the only available condition in practice should be a very small average error as long as training ends (Haykin, 1998).

Another issue is the ability to *generalize* well, meaning the network is adequately correct for test data never used in the training (Haykin, 1998). The poor *generalization* problem can be dealt with through an automated method, known as the *Cross-Validation* method. This can be used as an early stopping

method of training (Haykin, 1998). Ordinarily, a multilayer perceptron trained with the BP algorithm learns in stages, moving from the realization of fairly simple to more complex mapping functions as the training session processes. Overfitting can be identified through the use of cross-validation, for which the training data are split into an estimation subset and a validation subset. The estimation subset of examples is used to train the network in the usual way, except for a minor modification: the training session is stopped periodically (i.e. every so many epochs), and the network is tested on the validation subset after each period of training (Haykin, 1998).

2.2.3.2. Training methods. Three approaches were employed for training the ANN.

- *ANN-1* (*Simple training*)
  This is a fully connected model with two hidden layers and the number of neurons in each layer equal to three (the number of parameters). It was trained with the BP algorithm once.
- *ANN-2* (*Repeated training*)
  This was a fully connected model with two hidden layers and the number of neurons in each layer equal to three. It was trained with the BP algorithm 10 times. Each time the Mean Square Error (MSE) (Makridakis et al., 1998) over the training sample was calculated. The network with the smallest MSE error was used to produce the forecasts.
- *ANN-3* (*Advanced training*)
  This involved all different fully connected networks up to two hidden layers and no more than three neurons per hidden layer. The networks were trained 10 times each. The number of different networks was constrained so that the second hidden layer was only used as long as the first hidden layer contained three neurons. Each time the MSE over the training sample was calculated. The network with the smallest MSE was used to produce the forecasts.

2.2.4. Judgmental forecasts

A class of 43 senior engineering undergraduate students at the National Technical University of Athens, Greece (NTUA) were asked to produce forecasts of impact for the 12 hold-out TV programmes. To assist them in their task the students were provided with a table containing the values of the three cues (Importance, Competition and Time Zone) and the resulting impact for the 34 programmes that were broadcasting the sport events. (Note that subsequent forecasts were made by the same group as part of a study of the application of the Delphi method, but these results will not be discussed here.) A number of studies provide support for the use of students as proxies for professional forecasters and decision makers in experiments involving judgment (e.g. Remus, 1986).

3. The relative accuracy of the methods

The mean absolute error (MAE) was used to measure the forecasting accuracy of the methods. This was chosen in preference to measures based on absolute percentage errors (e.g. the MAPE and MdAPE) because these tend to be distorted by very small outcomes (Goodwin and Lawton, 1999) and because their main purpose is to allow forecast accuracy to be measured across series with different magnitudes and scales (Chatfield, 1988). The mean squared error (MSE) makes a stronger assumption than the MAE about the nature of the loss function that is associated with forecast errors and is more difficult to interpret. Its use has been criticised by Armstrong and Collopy (1992).

Table 3 shows the MAEs on the hold-out sample for the different forecasting methods. The methods are ranked in descending order of accuracy. A one-way repeated measures ANOVA indicated significant dif-

Table 3
Accuracy of the forecasting methods

Method	MAE
k-NN	8.90
R-pars	9.10
ANN-1	9.15
ANN-3	9.90
3-NN	12.68
Judgment	12.81
NN	12.89
R-step	14.40
R-all	14.60
ANN-2	16.15

ferences between the accuracy of the methods ($F_{11,88} = 2.68$, $p = 0.011$) (judgmental forecasts was excluded from this test because their results had a different structure – they consisted of 43 forecasts for each programme, rather than the single forecast obtained from the other methods). Duncan's post-hoc range test indicated both that *k*-NN, ANN-1 and *R*-pars were significantly more accurate than *R*-all and at least close to being significantly more accurate than *R*-step (e.g. the *p*-value between *R*-step and *R*-pars was 0.055). A separate *z*-test was used to determine whether the mean of the 43 judgmental forecasters' MAEs was significantly less than that of *R*-step This indicated the judgmental forecasters were, on average, more accurate than *R*-step ($z = 4.54$, $p < 0.0001$). ANN-2 fared particularly badly – it was significantly less accurate than all of the top five methods in Table 3.

To investigate the performance of the methods in greater depth two main aspects of the forecasts were examined: (i) the extent to which they suffered from systematic biases and (ii) how adept they were in utilising the cue information.

Table 4 gives details of the mean error (ME) of the methods (they are displayed in order of accuracy) and shows that they all tended to overestimate impact. However, this tendency was only very slight in the case of ANN-3, but very severe in the case of *R*-step, *R*-all and ANN-2. It can be seen that the higher levels of bias tend to be associated with the less accurate methods, but to investigate the extent to which inaccuracy is accounted for by bias, Theil's decomposition was applied to the forecasts (Theil, 1971). Theil's decomposition actually relates to MSEs, rather than MAEs but the two measures are closely related. The decomposition shows the contribution to the MSE of: (i) mean bias (i.e. the tendency to forecast to high or too low), (ii) regression bias (i.e. a tendency to forecast too high when the outcome is low and too low when the

Table 4
Mean error of the forecasting methods

Method	ME
k-NN	3.31
R-pars	−6.53
ANN-1	−1.90
ANN-3	−0.87
3-NN	−4.12
Judgment	−6.39
NN	−5.98
R-step	−10.46
R-all	−10.72
ANN-2	−14.22

Table 5
Extent to which biases contribute to MSEs

Method	Mean bias %	Regression bias %	Other causes %	Total
k-NN	8.7	8.0	83.3	100.0
R-pars	31.4	2.3	66.3	100.0
ANN-1	2.6	25.4	72.0	100.0
ANN-3	0.4	23.7	75.8	100.0
3-NN	6.8	37.0	56.2	100.0
Judgment	26.6	7.6	65.7	100.0
NN	11.5	41.7	46.8	100.0
R-step	33.2	29.9	37.0	100.0
R-all	33.0	26.3	40.7	100.0
ANN-2	49.5	17.7	32.7	100.0

outcome is high, or vice versa) and (iii) forecast error which is not caused by either of these biases (random error). The measurement of these three components is shown below:

$$\underset{\text{Mean bias}}{(\overline{A}-\overline{F})^2} + \underset{\text{Regression bias}}{(S_F - rS_A)^2} + \underset{\text{Random error}}{(1-r^2)S_A^2}$$

where $\overline{A}$ and $\overline{F}$ are the means of the actuals (A_t) and forecasts (F_t); S_A and S_F are the standard deviations of the actuals and forecasts; r is the correlation between the actuals and forecasts.

Table 5 shows the decomposition for each of the methods. It can be seen that bias is a major source of inaccuracy for the less successful methods, but in general other causes tend to predominate.

These other causes can be either inefficient use of the cue information or noise. The Brunswik lens model provides one framework for assessing the efficiency with which forecasting methods make use of cue information (Brunswik, 1952; Hammond, 1955). Designed as a method for assessing the quality of human judgment (Stewart and Lusk, 1994), the approach would also seem to be appropriate for assessing the output of computer-based methods that are based on the parallel architecture of the human brain, such as neural networks or methods which may directly operate in ways that are similar to the pattern matching tendencies of human judgment (Hoch and Schkade, 1996; Goodwin and Fildes, 1999), such as nearest neighbour analysis. The framework works by decomposing the correlation between the forecasts and outcomes into components which represent different aspects of cue use. Although the benchmark is an ideal *linear* model the decomposition also allows the analyst to assess the extent to which the alternative methods have detected non-linearities in the relationship. For our purposes the interesting components are:

(i) the predictability of the Impact of the TV programmes, given the cue information,
(ii) the extent to which a forecasting method's use of the cues matched that of an ideal linear model,
(iii) the consistency with which a method used the cue information, and
(iv) the extent to which the method was able to detect and use non-linearities in the relation between Impact and the cues.

3.1. Predictability

The lens model approach first provides an assessment of the extent to which the dependent variable is capable of being predicted by a *linear* model based on the cues. To obtain this measure a multiple regression model was fitted to the 12 out-of-sample outcomes and the cue values that applied for these events. The

multiple correlation coefficient associated with this model provides the measure of "predictability". The results are set out below:

The ideal *linear* model for the 12 hold-out sample results is:

$$\text{Impact} = 31.73 + 1.83\ \text{Importance} - 9.62\ \text{Competition} + 3.28\ \text{Time Zone}$$
$$R^2 = 52.8\%$$

The multiple correlation coefficient for this model (i.e. the level of "predictability" of Impact) is 0.726.

3.2. Match with ideal linear model

To determine the extent to which the forecasting methods matched the ideal linear model, multiple linear regression models are estimated by regressing each method's forecasts against the cues. For example, the mean forecasts obtained from the judgmental forecasters could be modelled as:

$$\text{Mean judgmental forecast} = 46.9 + 3.6\ \text{Importance} - 9.79\ \text{Competition} - 2.56\ \text{Time Zone}$$

Note that this corresponds closely to that of the ideal linear model. Not surprisingly, therefore, the correlation between the mean judgmental forecasts and those of the ideal linear model was 0.892. Table 6 shows the correlations for all of the methods in column 2. Many of the methods have forecasts that have high correlations with the ideal linear model. This is particular the case for *R*-pars. Note, however, that the forecasts of the *R*-all and *R*-step models only have low correlations with the ideal model. This may be because these models were overfitted to the in-sample data and therefore did not generalize well to the out of sample conditions.

3.3. Consistency

The next column of Table 6 displays the consistency with which the methods used the cues. This is measured by the correlation between the actual forecast they produced and the forecasts suggested by models of their cue utilization. Naturally, all three regression models had the maximum consistency of 1 since they must produce the same forecast for a given set of cue values. Surprisingly the judgmental forecasters have a high consistency of 0.882. Consistency is not an attribute normally associated with judgmental forecasting (e.g. O'Connor et al., 1993), but the forecasts modeled here are, of course, mean forecasts taken from a group of people so the averaging process may have removed most of the incon-

Table 6
How the methods used the cues

Method	Correlation with ideal linear model	Consistency	Detection of non-linearities
k-NN	0.767	0.991	0.785
R-pars	0.924	1.000	na
ANN-1	0.718	0.931	0.701
ANN-3	0.763	0.937	0.762
3-NN	0.654	0.914	0.672
Judgment	0.892	0.953	0.866
NN	0.811	0.848	0.747
R-step	0.521	1.000	na
R-all	0.505	1.000	na
ANN-2	0.537	0.970	0.606

na = not applicable.

sistencies associated with the individual forecasts. Of course, all of the non-judgmental methods applied their *rules* consistently, but their use of cue information may not have been consistent. For example, in the NN approach cue values of 3, 3 and 3 would have led to a forecast of 17% impact based on the one observation with this combination of values, but values of 2, 3 and 3 would also have led to a forecast of 17%. In other cases, a reduction in the first cue value from 3 to 2 would have reduced the impact prediction.

3.4. Detection of non-linearities

The final column of Table 6 displays estimates of the extent to which the methods were able to detect non-linearities in the relation between the cues and the outcomes. To obtain these estimates it was necessary to make further use of the multiple *linear* regression models for each method which related its forecasts to the available cue information. For example, the forecasts of the NN method could be modelled as:

$$\text{Forecast} = 69.70 - 0.23\ \text{Importance} - 14.5\ \text{Competition} - 3.84\ \text{Time Zone}$$

Now suppose we have a television programme for which the model of the NN method has a large positive residual, indicating that the forecast was much higher than would have been expected from the linear model of its cue use. Suppose that, for this programme, the *ideal* linear model also has a large positive residual–indicating that it substantially underestimated Impact. Thus we have a situation where the NN method produces a higher forecast than would be expected if it is making linear use of the cues at the same time as the ideal linear model is producing a forecast that is too low. This might imply that the NN method is making use of properties in the data that cannot be represented by a linear combination of cue values. In the case of judgment it might also imply that the forecaster is also able to use additional information that is not available to an "objective" model (e.g. that a pre-match "war of words" between the competing football coaches involved in a game is likely to have raised the level of interest in the event). We will refer to the ability of a forecasting method either to detect non-linear relationships between the cues and dependent variable or to use additional information as its ability to detect non-linearities.

A systematic tendency for a method to detect non-linearities can be detected by calculating the correlation between the residuals of the ideal linear model and the linear model of how each method used the cues. The last column of Table 6 suggests that *k*-NN, ANN-1 and ANN-3 were detecting non-linear relationships. Of course these methods only had access to information on the three provided cues. However, the judgmental forecasters also had access to a brief description of the sporting event which told them which teams were playing (i.e., qualitative information). This may, at least in part, account for the high correlation that appears in the last column of Table 6 for Judgment. In the case of the regression models, the detection of non-linearities is, by definition, not possible, so no entries have been included in the last column of the table for these methods.

4. Discussion

The key result from this study is that the *multiple* regression models performed relatively badly as a forecasting tool and were outperformed by either conceptually much simpler methods, like the bivariate (parsimonious) regression model, nearest neighbour analysis or by more complex methods like artificial neural networks (in certain forms). Forecasts based on human judgment also outperformed multiple regression.

First why did the nearest neighbour methods outperform multiple regression? After all, the best performing method of all, *k*-NN, based its forecasts on a sample of only three past observations, while the multiple regression forecasts were based on 34 observations. One explanation might be that the three independent

variables, Importance, Competition and Time Zone do not have separate additive effects on Impact. Instead, the effect on Impact may be derived through a complex interaction of the three variables so that it is highly dependent on specific combinations of values. These interactions may include complex if-then relationships. The *k*-NN method probably performed best of the nearest neighbour methods because its use of three cases would allow for some filtering of the noise associated with the observations. This appears to have more than compensated for by the fact that this method will need to include some cases that are less similar to the case being forecast than the single case that is used in NN.

Complex non-linear relationships can also be handled by artificial neural networks. Indeed, ANN models can effectively be viewed as complex regression models which include more terms, of an interactive and non-linear nature, than simple multiple linear regression models (although there are important differences between MLR and ANN in the model identification process and in the transparency of the resulting model (Fraser, 2000). The question is whether this extra complexity leads to accuracy gains. There was a clear distinction between the relative accuracy of ANN-1 and ANN-3 and the very poor performance of ANN-2. Recall that ANN-1 and ANN-2 have exactly the same structure and hence exactly the same ability to detect non-linearities. However, ANN-1 is trained only once while ANN-2 is trained 10 times (with MSE as a criterion). This means that ANN-2 is in danger of overfitting the in-sample data since it does not change structure and tries 10 times to achieve a better fit. Indeed, it is noticeable that the measures displayed in Tables 3–6 are very similar for ANN-2 and MLR suggesting that both methods suffered from overfitting. The relatively accurate performance of the bivariate regression model adds further support to this suggestion. The advantage of ANN-3 is its versatility in that it allows different structures to compete and so has less chance of overfitting and more potential to detect non-linearities.

Finally, why did human judgment, on average, outperform multiple regression? In applications of judgment to time series data, human forecasters often display a characteristic similar to overfitting in that they see imaginary systematic patterns in noise and attempt to extrapolate these patterns when producing their forecasts (Harvey, 1995). This tendency may be particularly associated with series that are presented graphically (Harvey and Bolger, 1996). However, in this study it seems unlikely that the judges would have the capacity to study simultaneously all the 34 past cases that were presented in a tabular format in a search for overall patterns. Given the limited information processing capacity of the human mind it seems more likely that they would search for a single similar case in a manner analogous to NN. This pattern matching behaviour has been observed in other studies of judgmental forecasters (e.g. Goodwin and Fildes, 1999). This idea is supported by the very close similarity between NN and Judgment in almost all of the measures presented in Tables 3–6. Given that the data here seemed to lend itself to reliable forecasting by nearest neighbour approaches then judgmental forecasts would be expected to be relatively accurate. Another factor that would favour judgment is the relatively small number of cues that the task involved. When forecasting tasks involve large numbers of cues, or large volumes of information, people resort to the use of heuristics to try to make the task manageable. These heuristics often lead to systematic biases in forecasts (Goodwin and Wright, 1993).

As mentioned earlier, the judgmental forecasters also had the potential advantage that, unlike the other methods, they could take into account the qualitative "event description" that was also provided. While the objective methods would each always produce the same forecast for a given set of values for the independent variables the judgmental forecasters could vary their forecasts by incorporating the estimated effect of this qualitative variable.

There are other interesting potential differences between the judgmental forecasters and the objective methods. First, it is unlikely that the names, or labels, given to the cues would be interpreted neutrally by the judgmental forecasters. There is plenty of evidence that judgmental forecasters are influenced by labels and the meanings that they seem to imply (Goodwin et al., 2004). Indeed labels create expectations about the statistical structure and, if this implied structure is congruent with the actual statistical structure, then this tends to improve performance (Sniezek, 1986). Congruent labels also tend to improve

the consistency of judgmental forecasts because they reduce the need to search for and test a large variety of hypotheses about the nature of the data. Second, the cue values here were assessments provided by an expert. While these values would have been treated unquestionably by the objective methods the judgmental forecasters would have had the opportunity to ignore or amend these values if they thought that they were not appropriate for the event that was being described.

Finally, it should be noted that, unlike MLR, there are no theoretical formulations for interval forecasts for nearest neighbour methods or for artificial neural networks. Usually these are calculated empirically based on the accuracy of the provided forecasts. There is also evidence that judgmental interval forecasts tend to be poorly calibrated (e.g. O'Connor and Lawrence, 1989). This may limit the usefulness of these alternative forecasting methods when decisions makers require information on the level of uncertainty associated with point forecasts.

5. Conclusions

The study has found that in a situation where MLR would be likely to be a popular choice of forecasting method, it was less accurate than a variety of other methods including those that were either more or less complex both conceptually and computationally. The relatively poor performance of MLR appears to result both from its tendency to over fit in-sample data and its inability to handle complex non-linearities in the data.

Of course, the MLR method was applied in an "automatic" mode here to make it comparable to the other objective forecasting methods examined in the study. In practice, its application might itself involve significant use of judgment. For example, judgment may be required in the examination of residuals to assess the validity of the underlying assumptions of MLR and in the interpretation of statistics, like t-tests. Decisions on whether to remove any outliers or influential data points will also be dependent on judgment. All of this would contribute to the relative cost and effort involved in using the method, while the potential extra benefits would be dependent on the expertise of the person applying the method and their ability to avoid judgmental biases during the modelling process.

Increases in computing power mean that methods like ANN and NN are now relatively quick and easy to apply and their application arguably involves less judgment than MLR. Where forecasts (as opposed to explanations of the behaviour of a system) are required, this study suggests that they should be considered as possible alternatives to MLR at an early stage and that they should receive greater prominence in standard courses in business forecasting and in associated textbooks.

Acknowledgements

The Lancaster University Management School (LUMS) Pump priming fund 2004/5 funded the work for this paper as part of a one year interdisciplinary project on the: "Development of an innovative forecasting methodology for television ratings". The project was started in April 2004, was supervised by Dr. K. Nikolopoulos and Professor Robert Fildes (LUMS). AGB Hellas S.A. made available the data for this study. AGB Hellas S.A. and the Institute of Communication and Computer Systems (ICCS) of the School of Electrical and Computer Engineering (ECE) in National Technical university of Athens (NTUA), co-funded the initial stages of this study during a research project towards a Ph.D. dissertation from Dr. A. Patelis entitled "Television Ratings Forecasting Methodology and Information System" (Patelis, 1999; Patelis et al., 2003). The project, started in 1997, lasted three years and was supervised by Professor Vassilis Assimakopoulos (NTUA). Paul Goodwin's contribution to this paper was supported by EPSRC grant GR/60198/01.

References

Armstrong, J.S., Collopy, F., 1992. Error measures for generalizing about forecasting methods – empirical comparisons. International Journal of Forecasting 8, 69–80.

Astebro, T., Elhedhli, S., 2003. The effectiveness of simple decision heuristics: A case study of experts' forecasts of the commercial success of early-stage ventures. Available from: <http://ssrn.com/abstract=579003>.

Bates, J.M., Granger, C.W.J., 1969. The combination of forecasts. Operational Research Quarterly 20, 451–468.

Brabec, B., Meister, R., 2001. A nearest-neighbour model for regional avalanche forecasting. Annals of Glaciology 32, 130–134.

Brunswik, E., 1952. The Conceptual Framework of Psychology. University of Chicago Press, Chicago.

Burger, C.J.S.C., Dohnal, M., Kathrada, M., Law, R., 2001. A practitioners guide to time-series methods for tourism demand forecasting – a case study of Durban, South Africa. Tourism Management 22 (4), 403–409.

Chatfield, C., 1988. Apples, oranges and mean squared error. International Journal of Forecasting 4, 515–518.

Chatfield, C., 1993. Neural networks – forecasting breakthrough or passing fad. International Journal of Forecasting 9, 1–3.

Chu, C.W., Zhang, G.P., 2003. A comparative study of linear and nonlinear models for aggregate retail sales forecasting. International Journal of Production Economics 86 (3), 217–231.

Clemen, R.T., 1989. Combining forecasts: A review and annotated bibliography. International Journal of Forecasting 5, 559–583.

Dawes, R.M., Corrigan, B., 1974. Linear models in decision making. Psychological Bulletin 81, 95–106.

Fernandez-Rodrguez, F., Sosvilla-Rivero, S., Garca-Artiles, M.D., 1999. Dancing with bulls and bears: Nearest-neighbour forecasts for the Nikkei index. Japan and the World Economy 11, 395–413.

Fraser, C.M., 2000. Neural networks: Literature Review from a Statistical Perspective California State University, Hayward – Spring 2000 (accessed 15/07/2005). Available from: <http://www.sci.csuhayward.edu/statistics/Neural/cfprojnn.htm>.

Gigerenzer, G., Todd, P., ABC Research Group, 2000. Simple Heuristics that Make us Smart. Oxford University Press, Oxford.

Goodwin, P., 2000. Correct or combine: Mechanically integrating judgmental forecasts with statistical methods. International Journal of Forecasting 16, 261–275.

Goodwin, P., 2002. Integrating management judgment and statistical methods to improve short-term forecasts. Omega: International Journal of Management Science 30, 127–135.

Goodwin, P., Fildes, R., 1999. Judgmental forecasts of time series affected by special events: Does providing a statistical forecast improve accuracy. Journal of Behavioral Decision Making 12, 37–53.

Goodwin, P., Lawton, R., 1999. On the asymmetry of the symmetric MAPE. International Journal of Forecasting 15, 405–408.

Goodwin, P., Önkal-Atay, D., Thomson, M.E., Pollock, A.C., Macaulay, A., 2004. Feedback-labelling synergies in judgmental stock price forecasting. Decision Support Systems 37, 175–186.

Goodwin, P., Wright, G., 1993. Improving judgmental time series forecasting: A review of the guidance provided by research. International Journal of Forecasting 9, 147–161.

Green, K.C., 2002. Forecasting decisions in conflict situations: A comparison of game theory, role-playing, and unaided judgement. International Journal of Forecasting 18, 321–344.

Hammond, K.R., 1955. Probabilistic functioning and the clinical method. Psychological Review 62, 255–262.

Härdle, W., 1992. Applied Nonparametric Regression (Econometric Society Monographs). Cambridge University Press.

Harvey, N., 1995. Why are judgments less consistent in less predictable task situations. Organizational Behaviour and Human Decision Processes 63, 247–263.

Harvey, N., Bolger, F., 1996. Graphs versus tables: Effects of data presentation format on judgemental forecasting. International Journal of Forecasting 12, 119–137.

Haykin, S., 1998. Neural Networks: A Comprehensive Foundation (International Edition. Pearson US Imports & PHIPEs.

Hoch, S.J., Schkade, D.A., 1996. A psychological approach to decision support systems. Management Science 42, 51–64.

Hogarth, R., Karelaia, N., 2004. Ignoring information in binary choice with continuous variables: When is less 'more'? Available from: <http://ssrn.com/abstract=501804>.

Kanas, A., 2003. Non-linear forecasts of stock returns. Journal of Forecasting 22, 299–315.

Lee, J.K., Yum, C.S., 1998. Judgmental adjustment in time series forecasting using neural networks. Decision Support Systems 22, 135–154.

Makridakis, S., Wheelwright, S., Hyndman, R., 1998. Forecasting Methods and Applications, third ed. Wiley, New York.

Meade, N., 2002. A comparison of the accuracy of short term foreign exchange forecasting methods. International Journal of Forecasting 18, 67–83.

Mentzer, J.T., Bienstock, C.C., 1998. Sales Forecasting Management. Sage Publications, Thousand Oaks, CA.

Mulhern, F.J., Caprara, R.J., 1994. A nearest-neighbour model for forecasting market response. International Journal of Forecasting 10, 191–207.

Nikolopoulos, K., Assimakopoulos, V., 2003. Theta intelligent forecasting information system. Industrial Management and Data Systems 103 (9), 711–726.

O'Connor, M., Lawrence, M., 1989. An examination of the accuracy of judgemental confidence intervals in time series forecasting. Journal of Forecasting 8, 141–155.

O'Connor, M., Remus, W., Griggs, K., 1993. Judgemental forecasting in times of change. International Journal of Forecasting 9, 163–172.

Patelis, A., 1999. Television Ratings Forecasting Methodology and Information System, Ph.D. Dissertation (in Greek), National Technical University of Athens Library, Athens.

Patelis, A., Metaxiotis, K., Nikolopoulos, K., Assimakopoulos, V., 2003. FORTV: Decision support system for forecasting television viewership. Journal of Computer Information Systems 43 (4), 100–107.

Remus, W., 1986. An empirical test of the use of graduate students as surrogates for managers in experiments on business decision making. Journal of Business Research 14, 20–30.

Sadownik, R., Barbosa, E.P., 1999. Short-term forecasting of industrial electricity consumption in Brazil. Journal of Forecasting 18 (3), 215–224.

Shanteau, J., 1992. How much information does and expert use? Is it relevant. Acta Psychologica 81, 75–86.

Sniezek, J.A., 1986. The role of labels in cue probability learning tasks. Organizational Behaviour and Human Decision Processes 38, 141–161.

Stewart, T.R., Lusk, C.M., 1994. Seven components of judgmental forecasting skill: Implications for research and the improvement of forecasts. Journal of Forecasting 13, 579–599.

Sun, H.Y., Liu, H.X., Xiao, H., He, R.R., Ran, B., 2003. Use of local linear regression model for short-term traffic forecasting. Initiatives in Information Technology and Geospatial Science for Transportation Transportation Research Record 1836, 143–150.

Theil, H., 1971. Applied Economic Forecasting. North Holland, Amsterdam.