Journal of Forecasting
J. Forecast. **19**, 515–535 (2000)

Evidence for the Selection of Forecasting Methods

NIGEL MEADE*
Imperial College, UK

ABSTRACT

Reid (1972) was among the first to argue that the relative accuracy of forecasting methods changes according to the properties of the time series. Comparative analyses of forecasting performance such as the M-Competition tend to support this argument. The issue addressed here is the usefulness of statistics summarizing the data available in a time series in predicting the relative accuracy of different forecasting methods. Nine forecasting methods are described and the literature suggesting summary statistics for choice of forecasting method is summarized. Based on this literature and further argument a set of these statistics is proposed for the analysis. These statistics are used as explanatory variables in predicting the relative performance of the nine methods using a set of simulated time series with known properties. These results are evaluated on observed data sets, the M-Competition data and Fildes Telecommunications data. The general conclusion is that the summary statistics can be used to select a good forecasting method (or set of methods) but not necessarily the best.

KEY WORDS ARIMA; ARARMA; Holt's linear trend; Holt's damped trend; robust trend; model selection

INTRODUCTION

The analysis of large sets of time series by Newbold and Granger (1974), Makridakis and Hibon (1979), Makridakis *et al.* (1982), Fildes (1992) and Fildes *et al.* (1997) has led to a number of conclusions about the relative merits of forecasting approaches and measures of forecasting performance. Of these conclusions, two are of particular relevance to the analysis to be described:

- The characteristics of the data series are an important factor in determining relative performance between methods.
- Statistically sophisticated or complex methods do not necessarily produce more accurate forecasts than simpler ones.

The issue of interest here is how much information is contained in the part of the time series used for model fitting, the fitted region, to suggest which forecasting method or class of methzods

* Correspondence to: Nigel Meade, The Management School, Imperial College, London SW7 2PG, UK. E-mail: n.meade@ic.ac.uk

Received January 1998
Accepted May 1999

will be superior. The distinction made between a model and a method is as follows. A model is an equation or set of equations representing the stochastic structure of the time series. A method is the combination of an estimation procedure and a model, possibly preceded by a model-identification procedure. The argument that there is information which can be used for the choice of method was put forward by Reid (1972). Based on his empirical experience, he produced a decision tree which guided the user to the most appropriate method, given the nature of the data. Jenkins (1974) argued that it was better to use the Box–Jenkins method to identify and estimate a model from the ARIMA class of time series models. This class includes models used by other forecasting methods as special cases. This view disagrees with both conclusions. In support of the first conclusion, Shah (1997) used discriminant analysis on a subset of 203 of the M-Competition time series. He demonstrated that a choice of forecasting method using summary statistics for an individual series was more accurate than using any single method for all the series considered. However, the choice was from only three methods, single exponential smoothing, Holt's linear trend and estimation of Harvey's basic structural model.

In the experiment described here, a set of time series is generated. Each time series has known properties and thus the most appropriate forecasting method is also known (i.e. the simplest method consistent with the time series model). For each series, data during the estimation period are summarized by a set of descriptive statistics. The objective is to discover the extent to which these statistics are useful in predicting which forecasting method will perform better. The predictive ability of these statistics is evaluated over a previously unused set of simulated time series, the series from the M-Competition and Fildes Telecommunications data.

The issues addressed here are important in the context of operations management where large numbers of time series are being forecast. The potential benefits offered are forecasts that are, on average, more accurate than those from any single forecasting method.

The structure of the paper is as follows. The forecasting methods considered are described first, and these include naive methods, exponential smoothing methods and ARMA-based methods. The data-generation process is described, and this includes details of the time series models and the simulation procedure. The literature suggesting summary statistics to aid the choice of forecasting method is summarized. Based on this literature and further argument, a set of summary statistics is proposed for the analysis. These statistics are then used as explanatory variables in predicting a performance index for each method. The effectiveness of the prediction of a forecasting method's relative performance is evaluated for simulated data and the data sets mentioned above.

FORECASTING METHODS

With the exception of two of the naive methods, the forecasting methods described here have been applied to the data in the M-Competition, either when it was held or subsequently. The 1001 time series from the M-Competition (Makridakis *et al.*, 1982) and 261 time series from Fildes (1992) were used to validate the performance of the forecasting methods described, because of their status and availability as 'standard' data sets. The exponential smoothing methods were validated using the summary performance figures published.

A time series, X_t, is observed for $t = 1, 2, \ldots, T$. The forecast origin is T and forecasts up to a horizon L are required. The time series models implicit in the following methods are described in

the next section. The first three methods assume a minimal time series structure and are usually called naive:

(1) Long-run average: an average of previous observations is used as a forecast:

$$E(X_{T+L}|X_T, \ldots, X_1) = \hat{\Theta}_0 \text{ for } L \geqslant 1$$

where

$$\hat{\Theta}_0 = \frac{1}{T}\sum_{i=1}^{T} X_i$$

(2) No change: the last known observation is used as a forecast:

$$E(X_{T+L}|X_T, \ldots, X_1) = X_T \text{ for } L \geqslant 1$$

(3) Deterministic trend: an average of previous differences is used to estimate a global linear trend:

$$E(X_{T+L}|X_T, \ldots, X_1) = X_T + L \cdot \hat{\theta}_0 \text{ for } L \geqslant 1$$

where

$$\hat{\theta}_0 = \frac{1}{T-1}\sum_{i=2}^{T}(X_i - X_{i-1})$$

The second group of methods use exponential smoothing. These methods assume that a time series can be decomposed into systematic and random variation, where systematic variation is described in terms of a local level and, optionally, a local trend. A full description is given by Gardner and McKenzie (1985). The methods considered are:

(4) Single exponential smoothing (stochastic level, no trend).
(5) Holt's linear trend (stochastic level, stochastic linear trend).
(6) Holt's damped trend (as method (5) but with the trend damped by a geometric factor).

Gardner and McKenzie show that the models implicit in these methods are all special cases of ARIMA models. In addition, the three models are nested. Exponential smoothing methods are sensitive to the estimation procedure used and to the values used to initialize this procedure. The procedure used here corresponds to that advocated by Gardner and McKenzie. The performance of the single exponential smoothing method used was compared with that published in Makridakis *et al*. (1982); the performances of Holt's linear and damped trend methods were compared with tables in Gardner and McKenzie (1985). The summary statistics in all these cases showed that these implementations duplicated the performance of those in the literature.

(7) Robust trend. This is a non-parametric version of Holt's linear trend method. The median-based estimate of trend is designed to be uninfluenced by outliers. See Grambsch and Stahel (1990). The performance of the robust trend method agreed with that in Fildes *et al*. (1997).

The third group comprises the ARMA-based methods: These are the sophisticated methods mentioned in the Introduction. The methods here use the ARIMA and ARARMA modelling frameworks:

(8) Autoregressive Moving Average (ARMA) models describe stationary time series. One major difference between the two methods is the means by which the series is transformed to stationarity. The ARIMA methodology proposed by Box and Jenkins (1970) is a cycle of identification and estimation. This procedure was followed in the 111 series sample in the M-Competition. Subsequently, automated procedures were used (for example, see Hill and Fildes, 1984). The procedure adopted here is as follows. Since only non-seasonal data are used, a basic ARIMA(p,d,q) is identified and estimated. The most crucial step in the identification process is deciding the value of d, the transformation to stationarity. The value of d is found using the GPH method (see Geweke and Porter-Hudek, 1983). This method is designed for use in an ARFIMA model where d is not constrained to integer values (FI is fractional integration). Given a non-integer d, this is converted to an integer by the following rule:

$$\begin{array}{lll} \text{if} & d < 0.5 & \text{then} \quad d \Rightarrow 0 \\ \text{otherwise} & & \quad\;\; d \Rightarrow \text{int}(d + 0.5) \end{array}$$

where int(z) is the largest integer less than or equal to z.

The differenced data is tested for a non-zero mean and adjusted accordingly before an ARMA model is identified. All ARMA models up to $p = q = 4$ are tested and the model chosen according to AIC (Akaike's Information Criterion). The ARIMA method used here differs from other published approaches, so it is interesting to compare its performance with the others. Table I shows that the forecasting performance of this approach over the 111 series sample from the M-Competition is broadly similar to other ARIMA implementations.

(9) The ARARMA methodology proposed by Parzen (1982) was applied with the benefit of human judgement (as was the ARIMA methodology) in the M-Competition. The methodology used here was validated in Meade and Smith (1985) and automated for use in Fildes *et al*. (1997). For the transformation of the data to stationarity, Parzen preferred a long-memory AR filter to the 'harsher' differencing used in ARIMA. In addition, a different approach to the identification of the ARMA model is used. Table II shows a comparison between Parzen's ARARMA forecasts and the procedure used here, and the performance is broadly similar.

Table I. Performance of ARIMA methods on 111 series sample of the M-Competition data

Horizon	M-Competition		Forecast pro		ARIMA method used here	
	MAPE	MdAPE	MAPE	MdAPE	MAPE	MdAPE
1	10.3	5.3	8.6	3.1	8.2	3.6
6	17.1	8.8	19.3	9.3	18.0	12.1
12	16.4	8.6	17.3	7.7	14.8	9.6
18	34.2	16.4	39.0	16.9	22.0	16.3

Table II. Performance of ARARMA methods on 111 series sample of the M-Competition data

Horizon	M-Competition		ARARMA method used here	
	MAPE	MdAPE	MAPE	MdAPE
1	10.6	4.8	8.4	4.1
6	14.7	9.0	15.7	9.5
12	13.7	6.6	14.7	9.8
18	26.5	11.6	20.1	15.5

Table III. The data-generating processes used

Structure	Equation	Process name and/or forecasting method
ARIMA (0, 0, 0)	$X_t = \Theta_0 + a_t$	White noise/long-run average
ARIMA (0, 1, 0)	$(1 - B)X_t = a_t$	Random walk/no-change forecast
ARIMA (0, 1, 0)	$(1 - B)X_t = \theta_0 + a_t$	Deterministic trend
ARIMA (0, 1, 1)	$X_t = v_t + \varepsilon_t \quad \varepsilon_t \sim \mathrm{N}(0, \sigma^2)$ $v_t = v_{t-1} + \phi s_{t-1} + \eta_t \quad \eta_t \sim \mathrm{N}(0, \sigma_\eta^2)$	Stochastic level only; $\phi = 0$/single exponential
ARIMA (0, 2, 2)	$s_t = s_{t-1} + \gamma_t \quad \gamma_t \sim \mathrm{N}(0, \sigma_\gamma^2)$ where $\sigma^2 > \sigma_\eta^2$ and $\sigma^2 > \sigma_\gamma^2$	Stochastic level and stochastic trend; $\phi = 1$/Holt's linear trend or robust trend
ARIMA (1, 1, 2)		Damped linear trend; $0 < \phi < 1$ Holt's damped trend
ARIMA (p, d, q)	$(1 - B)^d(1 - \cdots \phi_p B^p)$ $X_t = (1 - \cdots \theta_q B^q)a_t$	ARIMA
ARARMA	$\Phi(B)(1 - \cdots \phi_p B^p)X_t = (1 - \cdots \theta_q B^q)a_t$ where $\Phi(B) = (1 - \phi_L B^L)$ or $\Phi(B) = (1 - \phi_1 B - \phi_2 B^2)$	ARARMA

THE TIME SERIES GENERATION PROCESS

Test data were generated to provide a 'clean' environment for the comparison of forecasting methods and the calculation of summary statistics. This environment ensures that the implicit assumption of extrapolative forecasting, that the data-generating process for the forecast region is the same as for the estimation region, is valid. In exercises with real data, such as the M-Competition, this assumption may be violated and 'contaminate' results. The possibly naive philosophy here is to refine the results using simulated data and as the final stage evaluate them using real data.

The data-generating process used has three basic stages: a no-memory process or white noise; a short-memory filter; and a long-memory filter. Varying the nature of the short- and long-term filters produces a range of increasingly complex time series structures. The ARMA(p,q) is used as the short-memory filter. The short-memory series may then be passed through an integrating filter to produce an ARIMA model or an autoregressive filter to produce an ARARMA model. The structures were chosen to be theoretically consistent with the set of widely cited univariate forecasting methods described in the previous section. Eight different models were used to generate time series, and they are identified in Table III. The least complex forecasting method capable of forecasting the resulting time series is also named. Although there are ARIMA

representations of the linear models associated with the three exponential smoothing methods, the linear model shown was used as the data-generation process for the relevant data sets.

The random number generator used was a routine (RNNOR) from the IMSL Maths Library which generates pseudorandom numbers from a standard normal (Gaussian) distribution using an inverse cumulative density function technique. A Kolmogorov–Smirnov test of the null hypothesis of a Gaussian random variable using 5000 generated observations showed no evidence of a departure from this hypothesis. In addition, examination of the time series via autocorrelation function and periodogram showed no evidence of departure from the white-noise hypothesis. The white-noise series is used as the basis for the generation of all the above models. The parameters of the models were sampled from uniform distributions. Fildes *et al*. (1997) demonstrated that many time series in the M-Competition and the telecommunications data sets contained outliers in the first differences of the data. In order to replicate this property of observed (rather than artificially generated) time series, outliers were included in the generation process, and details are given in the Appendix. For each time series generated, 200 observations were produced, and the first 50 were discarded to remove possible contamination due to starting conditions.

VARIABLES DESCRIBING THE DATA IN THE FITTED REGION

These variables are summary statistics of the time series available for model estimation, and they will be used to (try to) explain and predict performance differences between forecasting methods. The hypothesis that the fitted region data contains information useful for choice of method has been put forward several times. Reid (1972) proposed a decision tree leading to method choice. Five variables were used to allow a choice between ARIMA and four types of exponential smoothing. Collopy and Armstrong (1992) prepared rules for an expert systems approach to extrapolation, and identify 'features identified by rules' which are variables describing the data.

The variables chosen for this analysis are described below and include those suggested by the sources mentioned. Number of observations—T: the series is classified as short (by Reid) if there are 50 or fewer observations. A comparison between random variation and non-random variation is suggested by Reid. This comparison is implemented by calculating the following statistics. The time series, X_t, is used in three regressions, each offering a form of non-random (or systematic variation):

LIN:	X_t	is regressed against	t
AR:	X_t	is regressed against	$X_{t-1}, \ldots, X_{t-m}$
ARD:	∇X_t	is regressed against	$\nabla X_{t-1}, \ldots, \nabla X_{t-m'}$

where m is min(20, $T/4$) and m' is min(20, $(T-1)/4$). For each regression, the adjusted R^2 and the variance ratio:

$$VR = \frac{\left(\frac{S_{\text{Err}}}{n-1}\right)}{\left(\frac{S_{\text{Reg}}}{k}\right)}$$

are calculated. (Where there are n observations and k independent variables in the regression for which the sums of squares can be written: $S_{\text{Total}} = S_{\text{Reg}} + S_{\text{Err}}$.)

A coefficient of variation is suggested by Collopy and Armstrong, and this is implemented here as

$$CV = \frac{\hat{\sigma}_{\text{Err}}}{\bar{X}}$$

where $\hat{\sigma}_{\text{Err}}$ is calculated from the regression of X_t against t (LIN).

Four binary variables are used to summarize other features suggested by Collopy and Armstrong:

(1) Index(trend) = 1 if basic and recent trends are in different directions
= 0 otherwise.

The basic trend is defined by the gradient of the (LIN) regression, and the recent trend is based on a similar regression using the last six observations.

(2) Index(sig. trend) = 1 if basic trend is significantly different from zero
= 0 otherwise

(3) Index(extreme) = 1 if $X_T > 0.9\ \text{Max}(X_1, \ldots, X_{T-1})$
or if $X_T < 1.1\ \text{Min}(X_1, \ldots, X_{T-1})$
= 0 otherwise

(4) Index(run) = 1 if the period to period changes of the last six observations are all in the same direction
= 0 otherwise

The presence of discontinuities and outlying observations is considered important by both sources. The frequency of outliers in period to period changes was used in Fildes *et al*. (1997) as a means of describing the differences between the Fildes data and the M-Competition data (denoted here as %out).

The autocorrelation function plays a role in the identification of an appropriate ARMA model so some correlations were calculated as summary statistics. The magnitude of these values will give partial information about the presence or absence of a trend. These values may also indicate whether an ARIMA model coinciding with one used by an exponential smoothing method is a feasible representation of the data. The correlations used are:

(1) r_k = |Correlation(X_t, X_{t-k})|.
(2) $r\nabla_k$ = |Correlation$(\nabla X_t, \nabla X_{t-k})$|.
(3) $r\nabla_k^2$ = |Correlation$(\nabla^2 X_t, \nabla^2 X_{t-k})$| for $k = 1, 2, 3$.

Tashman and Kruk (1996) considered a variance analysis protocol, suggested by Gardner and McKenzie (1988), for the selection of exponential smoothing procedures. The variances of the series and first and second differences of the series are calculated. The transformation with minimum variance suggests the appropriate exponential smoothing method: if it is the untransformed series then single exponential smoothing is appropriate; if it is the first differenced series then Holt's damped exponential smoothing is appropriate; if it is second differencing then Holt's

linear trend is suggested. Three binary variables are used to indicate the minimum variance transformation:

$V0 = 1$ if $V(X_t)$	$<$	minimum$(V(\nabla X_t), V(\nabla^2 X_t))$	0 otherwise
$V1 = 1$ if $V(\nabla X_t)$	$<$	minimum$(V(X_t), V(\nabla^2 X_t))$	0 otherwise
$V2 = 1$ if $V(\nabla^2 X_t)$	$<$	minimum$(V(X_t), V(\nabla X_t))$	0 otherwise

Since it was felt likely that the interactions between the summary statistics contained useful extra information, the interactions are included as extra variables. The interactions are simply the pairwise products of the statistics identified above.

EXPERIMENTAL PROCEDURE

Two sets of time series are generated. The first, the base data set, is used to evaluate the informational content of the summary statistics for the selection of an effective forecasting method. The second set is used later for out-of-sample validation of the selection rules proposed. For each data set, the number of observations used for model estimation and calculating summary statistics is varied (10, 20, 30, 40, 50, 75, 100 and 150 observations for estimation). The base data set contains 4096 series, 64 series for eight estimation lengths for eight data-generating processes. The validation set contains 2048 series, 32 in each sub-category. For each series in the base data set, the 25 summary statistics are calculated and forecasts are prepared using each of the nine forecasting methods mentioned. Forecasting performance is evaluated over a six-period horizon, and the accuracy measure used is the mean absolute error (MAE). The reason for this choice is to give a summary measure of forecasts over a short to medium horizon. This measure does not give undue importance to large errors (as a root mean square error would). In addition, since simulated series can have zero or negative values, it avoids dependence on the average level of the series (which a mean absolute percentage error would not).

In order to evaluate the informational content of the summary statistics, the existence of a requirement for selection must be demonstrated. For example, if single exponential smoothing always outperformed other methods, then one would always choose this method. Since relative performance between forecasting methods is the focus of this study, a performance index of method j on series i was created:

$$v_{ij} = \frac{MAE_{ij}}{\min[MAE_{i1}, \ldots, MAE_{i9}]}$$

For every series i at least one, v_{ij}, is unity. All the series ($j = 1, \ldots, 4096$) were forecasted by each of the forecasting methods ($i = 1, \ldots, 9$) described and the values of the performance index v_{ij} computed. The distributions of this index are summarized by the median and the mean in Table IV. The distribution of v_{ij} is, by its definition, skewed, thus presenting the median and the mean summarizes evidence of typical behaviour and extreme behaviour respectively. The table shows that for each data-generating process (dgp), the best or second-best performing forecasting method (in terms of median and mean v_{ij}) is the least complex capable of representing the process. Generally the methods which involve a set of nested models, such as Holt's damped trend, and ARIMA and ARARMA, perform well for most dgps. Data generated in accordance with an ARARMA process is usually best forecast by this forecasting method since the median

Table IV. Median and mean values of performance indices for nine methods on eight groups of 512 series of the base data set and 256 of the validation set. (In each row, the best value is shown in **bold**, second best in ***bold italic***)

Data-generating process		Forecasting method								
		Long-run average	No change	Deterministic trend	Single exponential	Holt's linear trend	Holt's damped trend	Robust trend	ARIMA	ARARMA
White noise	Median	**1.04**	1.18	1.26	1.05	1.17	1.10	1.28	1.06	***1.05***
	Mean	**1.20**	2.23	2.43	***1.27***	1.86	1.42	2.42	1.34	1.39
Random walk	Median	2.41	**1.23**	***1.27***	1.34	1.50	1.38	1.29	1.41	1.47
	Mean	4.34	**1.85**	2.05	2.08	2.44	2.17	***1.98***	2.26	2.58
Deterministic trend	Median	11.33	1.91	***1.20***	2.06	1.45	1.58	**1.15**	1.89	1.88
	Mean	21.94	2.76	***1.85***	3.16	2.32	2.61	**1.84**	3.04	3.36
Stochastic level	Median	1.34	***1.20***	1.26	**1.19**	1.26	1.24	1.27	1.26	1.30
	Mean	2.31	1.76	1.90	**1.62**	1.91	***1.69***	1.91	1.73	1.90
Stochastic level and linear trend	Median	11.93	2.13	***1.39***	2.49	**1.33**	1.43	1.43	1.69	2.02
	Mean	33.81	4.36	2.26	4.55	**2.23**	2.31	***2.23***	2.43	3.09
Stochastic level and damped trend	Median	2.71	1.39	***1.35***	1.39	1.44	**1.34**	1.36	1.40	1.50
	Mean	15.49	2.43	2.03	2.39	2.12	**1.79**	2.05	***1.99***	2.35
ARIMA	Median	1.75	1.74	1.60	1.48	1.40	***1.31***	1.61	**1.25**	1.35
	Mean	41.58	5.96	3.30	5.75	2.47	***2.26***	3.47	**2.15**	3.15
ARARMA	Median	2.21	2.40	2.44	1.75	1.80	1.70	2.74	***1.37***	**1.00**
	Mean	9.04	5.49	5.47	4.76	4.56	4.42	8.67	***3.71***	**1.48**
Overall	Median	2.27	1.46	1.39	1.42	1.40	***1.34***	1.41	**1.34**	1.35
	Mean	16.21	3.35	2.66	3.20	2.49	***2.33***	3.07	**2.33**	2.41
Validation data (2048 obs.)	Median	2.22	1.45	***1.32***	1.43	1.39	1.33	1.39	1.34	**1.30**
	Mean	21.44	3.92	2.95	3.81	2.48	**2.39**	3.33	***2.49***	2.59

performance measure was unity. The problems other methods had with ARIMA- and ARARMA-generated data are indicated by higher mean values for v_{ij}. The robust trend method performed well for data with a deterministic trend and with a stochastic linear trend.

At the foot of Table IV, there are summaries of the performance indices for the two generated data sets, the base data and the validation data. Table V gives a similar summary for the M-Competition data and the Telecommunications data, and in addition the mean and median absolute percentage errors (APEs) are given. Note that the figures refer to an average of forecasting horizons from one to six periods ahead. Holt's damped trend is one of the more consistently accurate methods. Using the mean performance index as a criterion, it is second best for the base data, and best for the validation and M-Competition data but its performance is mediocre for the Telecommunications data. Although a robust trend is clearly most accurate for the Telecommunications data, a deterministic trend outperforms a robust trend for both summary measures for the other data sets. The performance of ARIMA and Holt's damped trend is very similar for the base and validation data, but appreciably different for the observed data sets. These comments serve to demonstrate that the selection of the best method is a real problem, and there is not a single best forecasting method.

A PREDICTIVE EQUATION FOR A FORECASTING METHOD'S PERFORMANCE INDEX

In order to determine the usefulness of the summary statistics in selecting an effective forecasting technique an objective has to be established and a means of reaching that objective chosen. A possible objective would be to predict which method would forecast a series best. This objective would lead to a binary variable being assigned to each series—method combination, method j achieved minimum mean absolute error for series i or it did not. A classification procedure such as discriminant analysis could then be used as a means of testing the predictive content of the summary statistics. Shah (1997) uses several forms of parametric and non-parametric discriminant analysis to choose between three forecasting methods. However, the idea of a best method is not really satisfactory, since it implies that all the other methods are equally bad. A method giving 1.01 times the minimum mean absolute error is obviously preferable to one giving 2.00 times the minimum. The approach eventually chosen was to estimate the performance index v_{ij}:

$$f(v)_{ij} = \beta_{0j} + \sum_{k=1}^{K} \beta_{kj} X_{ik} + \varepsilon_{ij}$$

where $f(v_{ij})$ is a suitable transformation of the index, X_{ik} is the value of the kth summary statistic for series i, and the βs are parameters to be estimated. The objective is the minimization of the errors, ε_{ij}. Ordinary least squares was chosen as an easy implementable objective function. The minimization of

$$\sum_{i=1} \sum_{j=1} \varepsilon_{ij}^2$$

is a linear regression problem which decomposes into separate regressions for each method. The method expected to be most effective for a particular series would be the one with the predicted

Table V. Median and mean values of performance indices (PI) and absolute percentage errors (APE) for nine methods on the 1001 M-Competition series and the 261 series of the Telecommunications data set. (In each row, the best value is shown in **bold**, second best in ***bold italic***)

		Forecasting method								
Data-generating process		Long-run average	No change[a]	Deterministic trend	Single exponential	Holt's linear trend	Holt's damped trend	Robust trend	ARIMA	ARARMA
M1001 data	Median (PI)	3.65	1.40	***1.28***	1.34	**1.24**	1.28	1.29	1.31	1.34
	Mean (PI)	8.70	2.30	1.74	2.27	1.83	**1.67**	1.75	1.83	***1.73***
	Median APE	22.90	7.62	***6.60***	7.38	6.65	**6.38**	6.66	6.77	6.82
	Mean APE	34.69	15.01	14.71	14.81	16.05	**13.89**	14.92	14.77	***14.55***
Telecomms data	Median (PI)	33.26	2.92	1.48	2.92	1.66	1.91	**1.34**	1.66	***1.45***
	Mean (PI)	38.47	3.37	2.21	3.39	2.96	2.57	**1.86**	2.53	***1.98***
	Median APE	31.44	2.53	1.56	2.58	1.86	1.96	**1.42**	1.74	***1.54***
	Mean APE	38.72	3.89	2.98	3.92	3.56	3.18	**2.46**	3.26	***2.61***

[a] These figures for the M1001 data are directly comparable with those of Makridakis *et al.* (1982). The median figure agrees exactly. The mean APE published is 14.4, but the authors note in the text that APEs over 1000 are ignored. There is one APE of 3508 (out of 6006 considered) included above, which explains the discrepancy.

Table VI. Proportion of variation in transformed performance index explained by summary statistics and interactions as explanatory variables

Method	Adjusted R^2	Standard error	No. of variables
Long-run average	0.74	0.71	53
No change	0.40	0.61	39
Deterministic trend	0.28	0.57	41
Single exponential	0.43	0.56	44
Holt's linear trend	0.21	0.59	36
Holt's damped trend	0.23	0.52	40
Robust trend	0.31	0.60	45
ARIMA	0.18	0.57	37
ARARMA	0.26	0.58	35

performance index closest to unity. A dependent variable with a minimum value of unity is unsuitable for linear least squares estimation, so a transformation that stretches the range of values is desirable. The transformation chosen is:

$$f(v_{ij}) = \ln(v_{ij} - \alpha)$$

Choosing α as 1 gives the whole real line as a range, but since an observed v_{ij} can take 1 as a value this leads to negative infinite values. Examination of the regressions for values of α showed that the R^2 statistic (the proportion of variation explained) averaged over all methods, and reached a maximum for $\alpha = 0.9$. The regressions were carried out for each method and the results are summarized in Table VI by the adjusted R^2 statistics and the standard errors. A forward stepwise regression procedure was used for estimation, and some variables were subsequently deleted to ensure that all the βs are significant at a 5% level. It can be seen that the variations in performance of the long-run average are best predicted by the summary statistics, with an adjusted R^2 of 0.74. The performance of the no-change and single exponential smoothing methods are next easiest to explain with an adjusted R^2 around 0.4. The performance of ARIMA is most difficult to explain with an adjusted R^2 of 0.18. Thus the more specific the underlying model of the forecasting method, the easier it is to predict its relative performance. Conversely, the more general the underlying model is, the more difficult it is to predict its performance relative to other methods. For each series the method expected to be the most effective is that with lowest estimated transformed performance index, $f(v_{ij})$.

THE RELATIVE CONTRIBUTIONS OF THE SUMMARY STATISTICS TO PERFORMANCE PREDICTION

The set of β coefficients generated from the nine regressions represents the information content of the summary statistics with regard to predicting relative forecasting performance. Unfortunately the large number of coefficients makes obtaining any insight directly from this source difficult. Two indirect approaches are presented here. The information is distilled into a frequency table showing how often a variable is found significant, either individually or as part of an interaction term, in Table VII. The variables that appeared most often were the number of observations and the percentage of outliers, and the percentage of outliers was the most frequently used individual

Table VII. Frequencies with which summary statistics occur in performance predicting regression equations

	Number of times summary statistic appears in the regression equations																							
			In an interaction term with statistic code no.																					
	Code	Alone	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	% frequency occurrence in all terms in all equations
No. of observations, T	1	3	3																						14.6
% outliers	2	8	4	2																					12.1
Correlation (X_t, X_{t-1})	3	1	4	1	0																				4.8
Correlation (X_t, X_{t-2})	4	0	0	0	0	1																			5.9
Correlation (X_t, X_{t-3})	5	3	0	1	1	0	4																		9.3
Correlation $(\nabla X_t, \nabla X_{t-1})$	6	0	4	3	2	1	0	2																	4.5
Correlation $(\nabla X_t, \nabla X_{t-2})$	7	2	1	2	0	1	1	0	2																7.0
Correlation $(\nabla X_t, \nabla X_{t-3})$	8	0	2	0	0	0	1	0	0	1															2.5
Correlation $(\nabla^2 X_t, \nabla^2 X_{t-1})$	9	0	6	0	0	1	0	2	1	0	2														3.7
Correlation $(\nabla^2 X_t, \nabla^2 X_{t-2})$	10	0	1	2	0	0	1	0	3	2	0	0													2.8
Correlation $(\nabla^2 X_t, \nabla^2 X_{t-3})$	11	0	1	1	0	1	0	0	0	1	0	0	1												2.0
R^2 (LIN)	12	5	3	3	1	1	3	1	1	1	0	3	1	2											8.7
R^2 (AR)	13	0	0	0	0	0	0	0	0	0	0	0	0	0	0										2.8
R^2 (ARD)	14	1	5	7	3	1	0	6	2	1	2	0	1	3	0	8									4.8
CV	15	0	1	0	0	3	1	0	2	1	0	1	1	4	0	3	2								2.2
Index (trend)	16	0	1	0	2	3	0	1	2	0	3	0	0	3	0	1	0								2.8
Index (extreme)	17	1	2	2	0	1	2	0	2	0	0	1	0	0	1	1	1	0							3.1
Index (run)	18	1	2	1	0	0	4	0	2	0	0	2	3	3	0	0	0	3	2						1.7
Index (sig. trend)	19	3	2	3	3	4	2	2	0	1	1	0	0	4	0	0	2	1	3	0					2.5
VR (LIN)	20	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	2	1	0	2	0			0.0
VR (AR)	21	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	2	0	0	0	0	0		1.1
VR (ARD)	22	0	0	3	0	0	0	0	2	0	0	1	0	0	1	0	0	0	1	0	0	0	0	0	0.3
V0	23	0	1	1	1	2	1	0	3	0	1	2	0	2	8	0	1	0	0	1	2	0	0	1	0.0
V1	24	0	2	3	0	0	1	0	1	1	1	0	0	3	0	2	1	1	0	3	2	0	2	0	0.0
V2	25	3	4	0	3	1	4	2	0	0	3	0	0	1	0	1	1	1	3	1	0	0	2	0	0.8

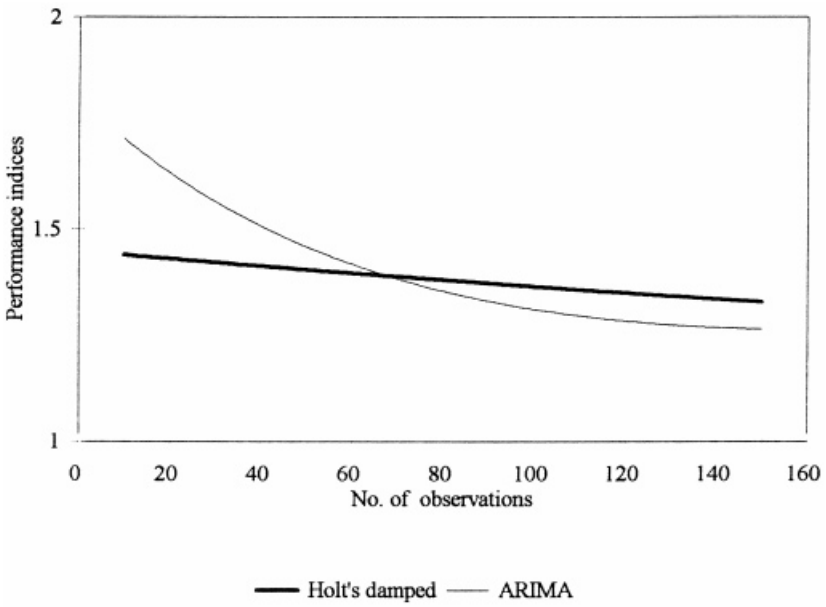

Figure 1. Performance Index for Holt's damped trend and ARIMA plotted against number of observations

variable, appearing in eight out of nine regression equations (it was not significant for the long-run average). The interaction term V0*R^2(AR) also appeared in eight out of nine equations, but it was not significant for the ARARMA equation. This term gives information about the proportion of variation explained by autoregressive terms for trendless series and is zero for a series with trend. In contrast, the variance ratio statistics were the least used group of variables.

Second, a selection of plots has been prepared to show the effect of some summary statistics on predicted performance. The median values of the 25 summary statistics for the base data are taken as default values for all measures not considered explicitly. The values of the interaction terms are adjusted appropriately. In Figure 1 the effect of the number of observations used for estimation on the relative performance of Holt's damped trend and ARIMA is shown. All else being equal, ARIMA is expected to perform better for a series of 70 observations or more. The effect of outliers on relative performance is shown in Figure 2. The performance of the robust trend is not appreciably effected by the proportion of outliers. In contrast, the performance of ARIMA and ARARMA deteriorates as the proportion of outliers increases. The summary statistic R^2(linear) is the proportion of variation explained by a linear trend. Figure 3 contrasts the performance of a no-change forecast with that of a robust trend as R^2(linear) increases. The performance of the latter improves as R^2(linear) increases, while that of the latter deteriorates.

It can be seen in Figure 4 that the performance of Holt's linear trend is not appreciably affected by R^2(linear). However, the performance of single exponential smoothing does deteriorate as R^2(linear) increases. If the value of index(run) is set to 1 rather than zero, the performance of single exponential smoothing deteriorates even faster. Index(run) is not a variable in the equation predicting the performance of Holt's linear trend. All four figures show plausible relationships but the figures are necessarily simplifications of a complex multivariate system. The plots are

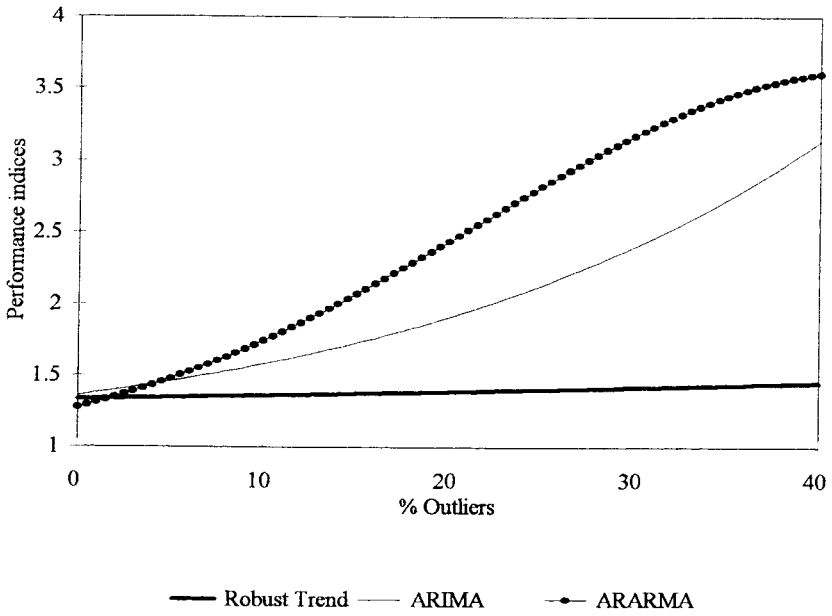

Figure 2. Performance Index for Robust trend ARIMA and ARARMA plotted against % outliers

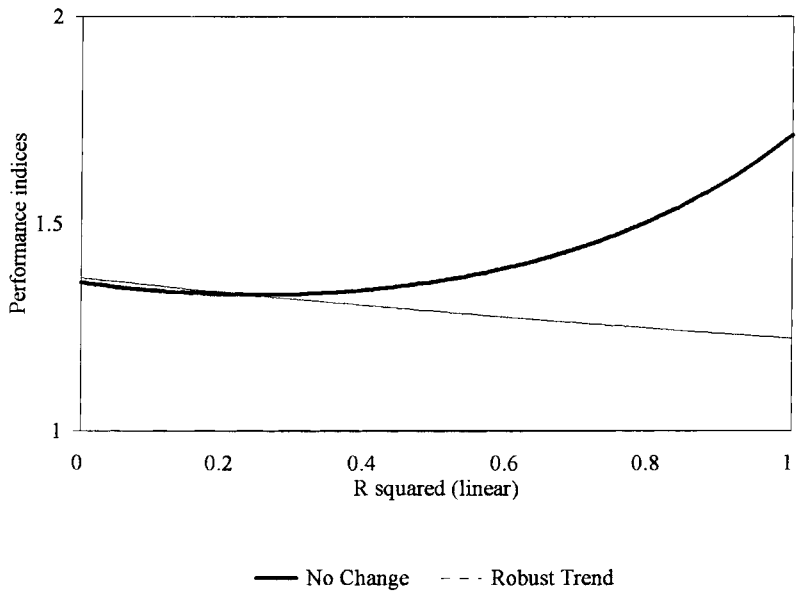

Figure 3. Performance Index for No change, Robust Trend and ARARMA plotted against *R* squared (linear)

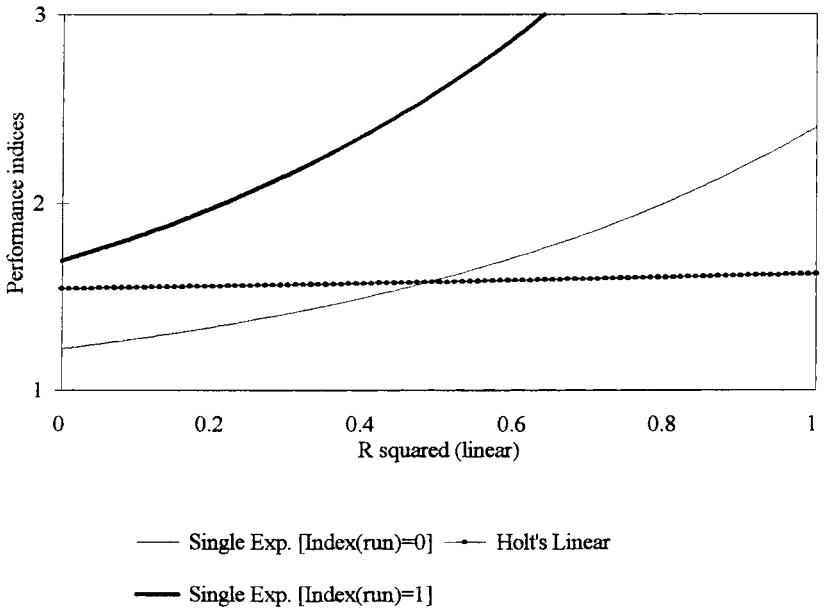

Figure 4. Performance Index for single exponential and Holt's linear trend plotted against R squared (linear) showing the effect of Index (run)

drawn keeping all other variables constant, ignoring the correlations between the summary statistics.

THE INFORMATION CONTENT OF THE PREDICTED PERFORMANCE INDEX

The values of the β coefficients estimated using the base data can be used in conjunction with the values of the summary statistics to predict the performance index of each forecasting method for a particular series. The usefulness of these predictions will be evaluated over the base data (an in sample analysis), the validation data, the M-Competition data and the Telecommunications data (out-of-sample analyses). Four evaluations are carried out:

(1) *Selection of the best naive method*: long-run average, no change and deterministic trend (see Table VIII). For the base and validation data, the method predicted to be most accurate does, in summary, outperform each of the naive methods. In fact, this informed selection of a naive model outperforms, in terms of the median, all the individual methods (see Table IV). For the M-Competition, the selection approach is ranked second; it fails to outperform its deterministic trend constituent. In contrast for the Telecommunications data, the selected naive method does outperform all of its constituents (see Table V).

(2) *Selection of the best linear trend model*: deterministic trend, Holt's linear trend and robust trend (see Table IX). Via the predicted performance from the regressions, the summary statistics provide sufficient information to provide a more accurate trend forecast than any single trend method for the base and validation data. For the M-Competition data, this is true for the mean and median performance measure. The median performance of the

Table VIII. Median and mean values of performance indices for the method selected from three naive methods

		Long-run average	No change	Deterministic trend	Method predicted to perform best out of three naive methods	
					Performance index	Rank (out of 4)
Base data (PI)	Median	2.27	1.46	1.39	1.23	1
	Mean	16.21	3.35	2.66	2.25	1
Validation data (PI)	Median	2.22	1.45	1.32	1.23	1
	Mean	21.44	3.92	2.95	2.78	1
M-Competition data (PI)	Median	3.65	1.40	1.28	1.32	2
	Mean	8.70	2.30	1.74	1.91	2
(APE)	Median	22.90	7.62	6.60	7.07	2
	Mean	34.69	15.01	14.71	15.01	2
Telecommunications data (PI)	Median	33.26	2.92	1.48	1.47	1
	Mean	38.47	3.37	2.21	2.19	1
(APE)	Median	31.44	2.53	1.56	1.56	2
	Mean	38.72	3.89	2.98	2.95	1

Table IX. Median and mean values of performance indices for the method selected from three linear trend methods

		Deterministic trend	Holt's linear trend	Robust trend	Method predicted to perform best out of three linear trend methods	
					Performance index	Rank (out of 4)
Base data (PI)	Median	1.39	1.40	1.41	1.26	1
	Mean	2.66	2.49	3.07	2.24	1
Validation data (PI)	Median	1.32	1.39	1.39	1.27	1
	Mean	2.95	2.48	3.33	2.28	1
M-Competition data (PI)	Median	1.28	1.24	1.29	1.24	1
	Mean	1.74	1.83	1.75	1.67	1
(APE)	Median	6.60	6.65	6.66	6.43	1
	Mean	14.71	16.05	14.92	15.01	3
Telecommunications data (PI)	Median	1.48	1.66	1.34	1.56	2
	Mean	2.21	2.96	1.86	2.40	3
(APE)	Median	1.56	1.86	1.42	1.69	3
	Mean	2.98	3.56	2.46	2.97	2

Table X. Median and mean values of performance indices for the single best method selected and for the combined forecasts of all methods with predicted indices of 1.4 or less

		Selected best method		Combined methods	
		Performance index	Rank (out of 10)	Performance index	Rank (out of 10)
Base data	Median	1.14	1	1.20	1
	Mean	1.66	1	1.64	1
Validation data	Median	1.17	1	1.21	1
	Mean	1.80	1	1.77	1
M-Competition data (PI)	Median	1.30	5	1.28	4
	Mean	1.70	2	1.65	1
(APE)	Median	6.67	5	6.36	1
	Mean	14.06	2	13.72	1
Telecommunications data (PI)	Median	1.60	4	1.63	4
	Mean	2.29	4	2.29	4
(APE)	Median	1.68	4	1.67	4
	Mean	2.93	3	2.92	3

selected method is marginally better than that of Holt's linear trend. This performance is carried over to the median APE but not the mean APE, where the selected method falls behind the deterministic trend and robust trends. For the Telecommunications data, characterized by strong negative trends, the selected method comes only second for the median and third for the mean performance index.

(3) *Selection of the best method from all methods* (see Table X). For the base data and the validation data, choosing the method with the best predicted performance leads to substantial improvements over the best individual methods for both the median and mean criteria. For the M-Competition data, the selected method is ranked fifth for median performance and second for mean performance. This performance is mirrored in the APEs. For the Telecommunications data, the selected method comes fourth for both criteria. This performance is better than ARIMA but less good than ARARMA, the deterministic trend and the robust trend method designed for this data set.

(4) *Selecting all methods with satisfactory predicted performance indices and forming an equally weighted combination of their forecasts* (see Table X). An alternative approach to selecting the method with the best predicted performance is the selection of a group of methods with better predicted performance. This approach was implemented by choosing a maximum predicted performance level. The resultant forecast is an equally weighted combination of forecasts of the methods with predicted performances below that level. The value of this threshold was optimized for the mean performance index of the combined forecast using the time series of the base data set. The value that minimized this objective was 1.4. For the base data and validation data sets, this combined forecast gives a marginal improvement in mean performance index at the expense of a deterioration in the median performance index. For the M-Competition data, the combined forecasts have the lowest mean performance index. The combined forecasts are ranked fourth in terms of median performance index behind Holt's linear trend, the deterministic trend and Holt's damped trend. However, in terms of APE, the combined forecasts yield the lowest mean and median values, slightly more

accurate than Holt's damped trend. For the Telecommunications data, the performance of the combined forecast is little different from the single best method.

SUMMARY AND CONCLUSIONS

To explore ways of identifying the more accurate forecasting methods for a time series, the data available in the estimation region has been summarized by a set of 25 statistics. The information content of these statistics has been evaluated using 'clean' simulated data sets and observed data sets, the M-Competition data and the Telecommunications data. The identification approach uses regression to effectively score each method for a given time series.

There are three major components to this analysis:

- The generation of the data.
- The choice of summary statistics.
- The method of performance prediction.

Each of these components could have been approached differently and the effect of choice of these components on the above conclusions is now discussed.

The generation of the data used the models underlying the forecasting methods and was thus driven by theoretical models, but this is virtually unavoidable in simulation. Alternatives include sampling real, observed data. The penalty for following this route is that the ability to generalize is lost because the results are biased by the data set used. The choice of summary statistics is extensive and within the group of 25 chosen there is much duplication of the information contained. Consequently, it may be asserted that a different choice of summary statistics would not lead to a substantial difference in conclusions, unless a new statistic is suggested which is effectively independent of those used. The performance prediction method used has some deficiencies, and the logarithmic transformation of the performance index as the dependent variable is not aesthetically pleasing. However, this method does overcome the deficiencies mentioned in the use of discriminant analysis. By using interaction terms some of the non-linear effects, that may be detected by a neural network approach, have been captured, but without the loss of transparency that the use of neural networks tends to involve.

This analysis suggests that the summary statistics do contain sufficient information to select a method or set of methods that will perform well. The use of the summary statistics allows an informed choice between naive methods representing broad classes of model (namely a long-run average, no change or a deterministic trend process). The evidence also suggests that the summary statistics can improve identification of an appropriate linear trend forecasting method from the choice of a deterministic, Holt's linear or a robust trend. Once the scope of method selection is widened, the comparison is between a method that envelopes a range of models (Holt's damped trend, ARIMA, ARARMA) and a method (or combination of methods) selected on the basis of the summary statistics. For well-behaved (in this case simulated) data sets, the selection outperforms any single method. For the M-Competition data, the selected combined forecasts perform well also. For the Telecommunications data, the selection process was unable to better the robust trend.

The argument in favour of using the summary statistics to select the forecasting method (or methods) is well illustrated if different prior experiences are imagined. If one approached the Telecommunications data with experience based on the M-Competition, one would expect Holt's

damped trend to perform well. The selected method was appreciably more accurate for the Telecommunications data than Holt's damped trend. Conversely, if the M-Competition data was approached with experience based on the Telecommunications data, one would expect the robust trend to perform well. In this case the combined forecast based on selection was more accurate than the robust trend.

APPENDIX: THE GENERATION OF OUTLIERS

After each time series was generated, a series of outliers, O_t, was generated and added to the series. The outliers could be either transient effects, ξ_t, or step effects, s_t, and these effects are combined as follows:

$$O_t = \xi_t + s_{t-1}$$

For each series a uniform (0, 1) random number, U_1, is sampled, and this is used to set the probability, p, that each observation is an outlier:

$$p = \begin{array}{ll} 0.015 & \text{if } U_1 < 0.25 \\ 0.055 & \text{otherwise} \end{array}$$

For each observation of the series, another uniform (0, 1) random number, U_2, is sampled, and this used to set the value of the transient term:

$$\xi_t = \begin{array}{ll} 0 & \text{if } U_2 \geqslant p \\ 6\sigma \ \text{sign}(Z_t)\max(|Z_t|, 2) & \text{otherwise} \end{array}$$

where Z_t is i.i.d. N(0, 1) and σ^2 is the variance of the noise term in the base series.

If $U_2 < p$, a third uniform (0, 1) random number, U_3, is sampled, and this is used to decide whether the outlier generates a step or a transient. Note that $s_1 = 0$:

$$s_{t+1} = \begin{array}{ll} s_t + \xi_t & \text{if } U_3 < 0.4 \\ s_t & \text{otherwise} \end{array}$$

This procedure for outlier generation is obviously arbitrary, and the effect of its application can be measured by measuring the proportion of outliers detected in the final series. This was done using the same procedure as Fildes *et al.* (1997). The quartile values for the generated data, the M-Competition data and the Telecommunications data are given in Table AI. As can be seen, the generated data exhibits broadly similar behaviour to the observed data sets.

Table AI. Percentage of outliers (in differenced series) detected in three data sets

	Base data	M-Competition data	Telecommunications data
Lower quartile	0.0	1.4	2.9
Median	6.1	4.1	5.7
Upper quartile	10.5	7.7	7.1

REFERENCES

Box, G. E. P. and Jenkins, G. M., *Time Series Analysis, Forecasting and Control*, San Francisco: Holden Day, 1970.

Collopy, F. and Armstrong, J. S., 'Rule-based forecasting: development and validation of an expert systems approach to combining time series extrapolations', *Management Science*, **38** (1992), 1394–1414.

Fildes, R., 'The evaluation of extrapolative forecasting methods (with discussion)', *International Journal of Forecasting*, **8** (1992), 81–111.

Fildes, R., Hibon, M., Makridakis, S. and Meade, N., 'The accuracy of extrapolative forecasting methods: additional empirical evidence', *International Journal of Forecasting*, **14** (1998), 339–58.

Gardner, E. S. Jr and McKenzie, E., 'Forecasting trends in time series', *Management Science*, **31** (1985), 1237–46.

Gardner, E. S. Jr and McKenzie, E., 'Model identification in exponential smoothing', *Journal of the Operational Research Society*, **39** (1988), 863–7.

Geweke, J. and Porter-Hudek, S., 'The estimation and application of long memory time series models', *Journal of Time Series Analysis*, **4** (1983), 221–38.

Grambsch, P. and Stahel, W. A., 'Forecasting demand for special services', *International Journal of Forecasting*, **6** (1990), 53–64.

Hill, G. and Fildes, R., 'The accuracy of extrapolation methods; an automatic Box–Jenkins package Sift', *Journal of Forecasting*, **3** (1984), 319–23.

Jenkins, G. M., 'Discussion of paper by Newbold and Granger', *Journal of the Royal Statistical Society, A*, **137** (1974), 148–50.

Makridakis, S. and Hibon, M., 'Accuracy of forecasting: an empirical investigation (with discussion)', *Journal of the Royal Statistical Society, A*, **142** (1979), 97–145.

Makridakis, S. *et al.*, 'The accuracy of extrapolation (time series) methods: results of a forecasting competition', *Journal of Forecasting*, **1** (1982), 111–53.

Meade, N. and Smith, I., 'ARARMA vs ARIMA—a study of the benefits of a new approach to forecasting', *Omega*, **13** (1985), 519–34.

Newbold, P. and Granger, C. W. J., 'Experience with forecasting univariate time series and the combination of forecasts (with discussion)', *Journal of the Royal Statistical Society, A*, **137** (1974), 131–65.

Parzen, E., 'ARARMA models for time series analysis and forecasting', *Journal of Forecasting*, **1** (1982), 67–82.

Reid, D. J., 'A comparison of forecasting techniques on economic time series', in Bramson, M. J., Helps, I. G. and Watson-Gandy, J. A. C. C. (eds), *Forecasting in Action*, Birmingham: OR Society, 1972.

Shah, C., 'Model selection in univariate time series forecasting using discriminant analysis', *International Journal of Forecasting*, **13** (1997), 489–500.

Tashman, L. J. and Kruk, J. M., 'The use of protocols to select exponential smoothing procedures: a reconsideration of forecasting competitions', *International Journal of Forecasting*, **12** (1996), 235–53.

Author's biography:

Nigel Meade is Reader in Management Science at the Management School, Imperial College, University of London. His research interests are statistical model building in general and applied time series analysis and forecasting in particular. He is currently Chairman of the UK OR Society Forecasting Study Group.

Author's address:

Nigel Meade, The Management School, Imperial College, London SW7 2PG, UK.

