Poler Escoto, R.; Mula, J. (2011). Forecasting model selection through out-of-sample rolling horizon weighted errors. Expert Systems with Applications. 38(12):14778-14785. doi:10.1016/j.eswa.2011.05.072.

Forecasting model selection through out-of-sample rolling horizon weighted errors

Raul Poler, Josefa Mula

CIGIP (Research Centre on Production Management and Engineering)
Universidad Politécnica de Valencia
EPSA – Pza Ferrándiz y Carbonell, 2
03801 Alcoy, Spain
Tel. +34 966528423 Fax +34 966528585
email: rpoler@cigip.upv.es

Abstract

Demand Forecasting is an essential process for any firm whether it is a supplier, manufacturer or retailer. A large number of research works about time series forecast techniques exists in the literature, and there are many time series forecasting tools. In many cases, however, selecting the best time series forecasting model for each time series to be dealt with is still a complex problem. In this paper, a new automatic selection procedure of time series forecasting models is proposed. The selection criterion has been tested using the set of monthly time series of the M3 Competition and two basic forecasting models obtaining interesting results. This selection criterion has been implemented in a forecasting expert system and applied to a real case, a firm that produces steel products for construction, which automatically performs monthly forecasts on tens of thousands of time series. As result, the firm has increased the level of success in its demand forecasts.

Keywords: Expert System, Forecasting model selection, Time series, Automatic forecasting, Error measures

1. Introduction

In the business setting, demand forecasting must be considered a process to obtain information that will be used in different decision-making processes. Success in demand forecasting is a critical factor so that the cost-cutting and improved customer service objectives in planning processes and production scheduling are met (Spedding & Chan, 2000). In short, reducing errors in forecasts helps minimise the risk that the firm assumes to cover its customers' demands (Nikolopoulos & Assimakopoulos, 2003).

In recent decades, numerous time series forecasting models have been proposed. A review of the last 25 years may be seen in De Gooijer & Hyndman (2005). Time series forecasting software tools usually offer a variety of techniques, some of which provide the user the possibility to automatically define parameters. In real business settings however, where it might be necessary to forecast thousands of time series, it is necessary to provide the decision-maker with expert systems that either deal with the automatic parameterisation of certain forecasting models or the most suitable forecasting model selection from a set of models.

On the other hand, the forecasting error concept has also been an object for various authors in the literature. This measurement is done with varied objectives: to determine the degree of success of a specific forecast, to analyse how forecasting evolves over time, to compare different forecasting models with a time series, to check the performance of a forecasting model with several time series, etc. In terms of the objective pursued,

measuring the error may be done in several ways: out-of-sample and within-sample errors, with a sign or in an absolute value, that are linear or quadric, with units or adimensional errors, etc. Besides, some forecasting models will be more suitable than others depending on the impact that the error made in the immediate, short-term, mid-term or long-term forecasts has on the business decisions.

The main objective of this work is to propose an automatic selection method of time series forecasting models which is applicable to different work settings and that allows the user to consider the importance of past errors.

The rest of the article is set out as follows: the measurements of forecasting errors are analysed in Section 2. In Section 3, different model selection criteria are analysed. Section 4 analyses the treatment of demand forecasting in expert systems and decision support systems. An extended criterion for time series forecasting models is proposed in Section 5. The results of applying the proposed selection criterion are described in Section 6 by using a two forecasting models on the monthly M3 Competition series, and its application to a real case is also described. Section 7 presents the conclusions and future lines of research are considered.

2. Forecasting errors

A wide range of formulae exists to measure forecasting errors. From basic formulae to those that use relative errors, which take either the time series values or the forecasts obtained by a forecasting model as a reference. The model which is used more often as reference is Naïve (the forecast equals the last value observed). In De Gooijer & Hyndman, 2005, the more commonly used forecasting error measurements are related.

The difference between the absolute and quadratic errors is that the latter penalise the more serious errors to a greater extent. The selection between both types will depend on the importance given to major forecasting errors. Theil's inequality coefficient represents a compromise between absolute and quadratic errors (Makridakis *et al.*, 1998).

It will be appropriate to use certain errors or others depending on the objective pursued. In order to compare the performance of a forecasting model on different time series for example, it is advisable to use adimensional errors. This is achieved simply by dividing by the time series value (MAPE). Nonetheless, dividing only by the time series value means that the error is not symmetrical (the relative error differs in terms of the error sign for identical errors in the absolute value). Nonetheless, the sMAPE may be used to avoid this problem. However, it has to be taken into account that the sMAPE presents an undesired performance when the real value or the forecast come close to zero. If a reference model is used (like Naïve) to obtained a relative error, at the same time a model goodness is obtained in relation to the reference model (which tends to be the most simple). Armstrong & Collopy (1992) recommended the Geometric Mean of the Relative Absolute Error (GMRAE) for model calibration. For selecting the most accurate methods they recommended the Median RAE (MdRAE) when few series are available and the Median Absolute Percentage Error (MdAPE) otherwise.

Strictly speaking, forecasting errors are observed as the difference between the forecast and the real value when the forecast was done prior to knowing such a value. Nevertheless, certain time series forecasting models are based on the determination of their parameters so

that the time series fits. In such a case, a 'forecast' of the already known real values is carried out. In this way, the difference lies between within-sample forecast for a forecast with already known values, and an out-of-sample forecast for a forecast that involves unknown values. The errors obtained in this way are named within-sample and out-of-sample, respectively.

Researchers generally agree forecast accuracy should be assessed using out-of-sample tests rather than in-sample tests (Fildes & Makridakis, 1995). For a given forecasting method, in-sample errors are likely to understate forecasting errors. The M-competition (Makridakis *et al.*, 1982) and other empirical studies show that forecasting errors generally exceed in-sample errors. Overfitting and structural changes may further aggravate the divergence between in-sample and post-sample performance (Tashman, 2000).

In short, it is considered that forecasting out-of sample errors is more suitable for comparing time series forecasting models because their good model fit to time series data does not have to imply high accuracy in future forecasts (Simkins, 1995).

Calculating out-of-sample forecast errors is done at the same time as the real value is known. Nonetheless, this forecasting may be done by what some authors call the out-of-sample simulation (Coccari & Galucci, 1984). This is done by simulating that the last real time series values are not known and by applying the forecasting model by calculating the out-of-sample error obtained. Makridakis (1990) used the sliding simulation as a process for method selection and estimation. Makridakis applied variants of the sliding simulation to some time series used in the M-competition (Makridakis *et al.*, 1982) and demonstrated that post-sample forecasting accuracy improved when smoothing weights were calibrated to minimize the post-sample error instead of calibrating weights in-sample. Fildes (1989) also used the rolling horizon procedure to compare the efficacy of various method-selection rules. Weiss & Anderson (1984) analysed the case of cumulative forecasts and proposed a calibration of the forecasting model to minimise a cumulative post-sample error measure.

The present study intends to determine which forecasting model, from those that form a given series, is the most appropriate for each time series analysed. Therefore, since the comparison among the various time series is not a requisite, the indicators used are not relative errors, thus avoiding undesired performances with extreme values.

3. Selection models methods

The first forecasting model selection proposals date back to the sixties, and centred on the model's goodness-of-fit to values that had already been observed. However, this selection approach prioritises models with more variables as they achieve a better fit, but a good ex-post forecast does not guarantee a good forecast of future values. Miller (1990) did a detailed study of the selection methods of those models proposed until that time.

The problem of finding the equilibrium between the goodness-of-fit and the model's complexity is ruled out if the approach changes and if the indicators related with out-of sample forecast errors are chosen. The situation might arise where a complex model with very good fits produces greater out-of-sample errors than the simpler model with a worse fit.

Generally, the two most used model selection criteria are those of Akaike (*Akaike Information Criterion* AIC) (Akaike, 1973) and of Schwarz (*Bayesian Information*

Criterion BIC) (Schwarz, 1978). Both contemplate the model's complexity by taking into account the degrees of freedom (the number of estimated parameters). The Schwarz criterion penalises the use of degrees of freedom more acutely (more complex models), thus it is stated to be more consistent. On the other hand, the Akaike criterion is asymptotically efficacious, while Schwarz's is not. Asymptotic efficacy is related to the hypothesis that reality is much more complex than any model considered because, when the number of observations is increased, the number of the models in the series considered must also increase. The criteria that become slower as the number of models increases are not asymptomatically efficient.

An alternative model selection method is the so-called cross validation method (Stone, 1974). Data is partitioned into subsets in such a way that the analysis is initially performed on a single subset, while the other subset(s) is/are retained for subsequent use in confirming and validating the initial analysis. The mean squared prediction errors in the subsets left out defines the cross validation error.

Inoue & Filian (2006) analyzed the Information Criteria (IC) against the simulated out-of-sample forecasts for model selection. They compared the asymptotic and finite-sample properties of these methods in terms of their ability to minimize the true out-of-sample prediction mean squared error. They showed that, under suitable conditions, the IC method will be consistent for the best approximating model among the candidate models. In contrast, under standard assumptions the simulated out-of-sample forecasts method, based on recursive or rolling regressions, will select overparameterized models with positive probability.

The relationship between model performance and certain characteristics of the time series has been a research topic for several authors (Makridakis *et al.*, 1982). Time series can be classified into subcategories (yearly, quarterly, and monthly data, micro and macro data, industry and demographic data, and seasonal and non seasonal data). However, the pattern must also be taken into account to obtain differences of performance between selected and non selected models.

There are two basic sources of knowledge about forecasting method selection: empirical studies and forecasting experts (Collopy & Armstrong, 1989). The empirical literature provides numerous guidelines for selecting among forecasting methods. Rule-based forecasting (RBF) is a type of expert system that is applied to time series extrapolation. The rules are based on the knowledge of five experts on forecasting methods. It consists of 99 rules and the forecast is obtained from combining forecasts from four extrapolation methods: random walk, regression, Brown's linear exponential smoothing trend (Brown, 1959) and Holt's exponential smoothing (Holt *et al.*, 1960), using 18 features of time series (Collopy & Armstrong, 1992) (Armstrong *et al.*, 2001).

Franses & Koehler (1998) proposed a model selection strategy for time series that displays increasing seasonal variation. This strategy provides a systematic overall approach without using Box-Cox transformations for comparing models with alternative stochastic and deterministic components. Their empirical results, however, indicate that the models selected using our tests on the in-sample observations often perform reasonably well in out-of-sample forecasting.

Machine learning algorithms can be applied to the selection of forecasting methods (Arinze, 1994) as a classification problem where the best forecasting model is the class attribute, and where time series features are the predictors. These algorithms learn to relate

time series features with the best models. Using such an approach, Prudencio *et al.* (2004) propose a supervised classification method that originates from the symbolic data analysis field for the model selection problem.

Flores & Pearce (2000) described an expert system which was originally designed to forecast monthly demand for industrial products that was modified to run the data of the M3 Competition. The rule base for this forecasting expert system was implemented using the IF-THEN rules. The complete process detects and adjusts irrelevant early data, detects and adjusts outliers, verifies the trend type, detects seasonality and period, chooses the preferred forecasting method, generates forecasts, shows forecasts from all methods to human users, allows users to choose a preferred method alters forecast values and finally stores the results. The forecasting methods used were: simple exponential smoothing, Gardner's damped trend exponential smoothing, classical decomposition (all of them without and with seasonality) and six-period moving average. In addition, a combination approach which averages the forecasts of all methods was used.

Most selection criteria of forecasting models based on out-of-sample errors use the one-step forecasting error, in other words, the errors obtained in the forecast of the next point of the time series. Nonetheless, and depending on how the firm uses immediate, short-, mid- and long-term forecasts, a more suitable forecasting model might be that with, for instance, a mid-term out-of-sample content than another model with a very good one-step out-of-sample content, but one with a poorer mid-term performance. In short, the importance of the forecasting horizon must be taken into account when defining a model selection criterion.

4. Forecasting Experts Sytems

Forecasting demand is a complex area of decision making in the company. Not only lot of variables have to been considered, but normally the forecasting process should be repeated on a wide range of products, making it necessary to provide an automatic system to the decision maker to perform this work. However, demand forecasting has been a decision process lowly implemented in expert systems, and in some cases, indicators of the goodness of the results have not been the most appropriated from a business standpoint. In the next paragraphs some of the most significant studies on these issues are discussed.

Wong and Chong (1993) performed a survey about applications of expert systems in manufacturing industries. The results of this study show that expert systems are generally not perceived to be accurate or reliable tools for forecasting production demand. The review of pertinent literature also revealed that few expert systems were available for forecasting in manufacturing. They concluded that intelligently combining forecasts obtained from different sources and forecasting techniques was a promising area of research.

Armstrong and Yokum (2001) performed a survey among forecasters to judge potential adoption of expert systems in relation to Box-Jenkins and scenarios. The respondents were classified in researchers, educators, practitioners and decision makers. The respondents viewed favourably expert systems in comparison with the mentioned techniques. They concluded that, in comparison with two well-established forecasting techniques, expert systems appear to have reasonable prospects for diffusion.

Liao (2005) performed a literature review and classification of articles from 1995 to 2004 about expert systems. The applications to forecasting techniques were related with Neural Networks and Fuzzy Expert Systems.

Eom and Kim (2006) conducted a follow-up survey about Decision Support Systems (DSS) Applications covering the period between 1995 and 2001 extending the two previous ones (Eom and Lee, 1990), (Eom *et al.*, 1998). They conclude that the production and operations management applications were the predominant DSS application area over the 90s and stated that the second survey reported several applications for aggregate demand or item demand forecasting and that the success of most industries often hinges on the accuracy of their forecast of demand. Forecasting and statistical models among others have been increasingly embedded in the model base of DSSs.

Researchers have developed expert systems or decision support systems for demand forecasting to face not only the time series forecasting, but to model complicated aspects as promotion, incomplete information or new product launching among others.

Kuo and Xue (1998) proposed a decision support system that utilizes fuzzy logic and a fuzzy neural network for the sake of learning fuzzy IF–THEN rules obtained from the marketing experts with respect to promotion, the result is further integrated with the forecast from artificial neural networks using the time series data. Their study concludes that the proposed system performs more accurately forecast than the conventional statistical method and single artificial neural networks.

Efendigi *et al.* (2009) proposed a forecasting mechanism modelled by artificial intelligence approaches by the comparison of both artificial neural networks and adaptive network-based fuzzy inference system techniques to manage the fuzzy demand with incomplete information. The bipartite methodology obtained more accurate forecasts and they considered their proposal as a successful decision support tool in forecasting customer demands.

Ching-Chin *et al.* (2010) proposed a decision-support system called the New Product Forecast System, to help execute the standard forecast procedure for new product sales forecasts. The Forecasting Model module contains the templates of six forecasting methods both classical (Moving Average, Exponential Smoothing, and Exponential Smoothing with Trends) and heuristic (Taylor Series, Sales Index, and Diffusion Model) specifically designed for new product sales forecasting. The decision support tool allows calculating the forecast according to the desired planning horizon: short-term, mid-term, and/or long-term.

In order to state the success of an expert system for demand forecasting, the main parameter to consider is the forecast accuracy, i.e. minimizing forecasting errors. In Section 2 different commonly used forecasting errors have been treated. Its use in expert system for demand forecasting is varied: arithmetic mean, mean absolute error, mean squared error, root mean square error or mean absolute percentage error.

Petrovic *et al.* (2006) proposed a decision support system for demand forecasting that combines four forecasts values: two of them represent subjective judgments on future demand and two crisp values obtained using conventional statistical methods. The learning mechanism consider the arithmetic mean of the forecasts errors recorded in the past periods.

Ali *et al.* (2009) used the mean absolute error (MAE) as a measure of overall accuracy of the candidate models on an extensive SKU-store level sales and promotion time series

from a European grocery retailer. They used regression trees with explicit features to model promotion and reported a substantial improvement of accuracy.

Efendigil *et al.* (2009) proposed a forecasting mechanism modelled by artificial intelligence approaches including the comparison of artificial neural networks and adaptive network-based fuzzy inference system techniques to manage fuzzy demand. They employed the mean squared error (MSE) and the mean absolute percentage error (MAPE) as training errors.

Lin and Lee (2009) combined Grey forecasting (GM) and Markov–Fourier Grey forecasting model (MFGM) to develop an expert system of diagnosis by artificial intelligence which improves the effectiveness of forecasting randomly fluctuating data. They used use the mean absolute error (MAE) to compare forecasting.

Pai *et al.* (2009) developed a seasonal support vector regression (SSVR) model to forecast seasonal time series data. To minimize forecasting error of the SVR model they use the negative values of the mean absolute percentage error (MAPE) and, in addition, the root mean square error (RMSE) is used to measure the forecasting accuracy.

Sayed *et al.* (2009) proposed an integrated model of statistical methods and a genetic algorithm used to choose the best weights among the statistical methods and to optimize the forecasted activities combinations that maximize profit. They consider the mean squared error (MSE) to guide the genetic algorithm for searching for the best combination of weights between the methods.

Singh (2009) presented a method of forecasting based on high-order fuzzy time series. The comparison of accuracy in forecasted values of the proposed models with other models is made on the basis of mean absolute percentage error (MAPE) and mean square error (MSE).

Behnamian and Ghomi (2010) introduced a new time-series forecasting model based on non linear regression with high flexibility to fit any number of data without preassumptions about real patterns of data and its fitness function. The proposed hybrid approach comprises two components: a particle swarm optimization and a simulated annealing. They use the mean absolute error (MAE), the mean square error (MSE) and the root mean square error (RMSE) but also the mean absolute percentage error (MAPE) in a case study.

Khashei and Bijari (2010) proposed a hybrid model of artificial neural networks using auto-regressive integrated moving average (ARIMA) models in order to yield a more accurate forecasting model than artificial neural networks. They used as cost function the mean squared error (MSE) for training and mean absolute error (MAE) and mean squared error (MSE) to compare with other forecasting methods.

In most of the expert systems one-step forecasting error is used to calculate the forecasting accuracy, but the real cases show that firms are interested in good forecasts not only at short-term but over a time horizon. This fact will be taken into account in the next section which describes the proposal for the automatic selection of forecasting models.

5. Proposal of a model selection method

Choosing one selection criterion of time series forecasting models will not only depend on the setting to which they are to be applied, but also on the various considerations to be

taken into account. The characteristics of the type of problem that we wish to solve are:

- The set of time series on which to proceed and, therefore, for which the most suitable forecasting model must be selected, is very large (more than 10,000).
- The group of time series forecasting models considered is wide and contains models of various kinds.
- The forecasts to be achieved are not only of the one-step kind, but of a wide horizon (18 periods, for instance).
- The decision-maker is interested in taking into account the errors made in the whole forecasting horizon and also at the different times the forecast was done. The decision-maker is also interested in taking decisions about the degree of power of major errors.
- It must be possible for the method to be applied automatically without the need of the decision-maker intervening.

For the purpose of facilitating the design of the selection criterion, several sentences are dedicated to help create their formula:

- The selection criteria of the models will operate with out-of-sample errors rather than within-sample errors.
- All errors made by all the forecasting models will be recorded for the forecasting horizon considered.
- A different importance will be given to the errors made throughout the forecasting horizon.
- A different importance will be given to the errors which, for the same period, have produced a specific forecasting model applied at different moments in time.
- A power could be defined for the errors which depends on the closeness of the period considered.

The following variables have been included:

T : current period (last real observation)

P : periods in which forecasts are calculated

Y_t : the real time series value in period t

F_t^s : forecast of period t calculated in period s

$e_t^s = Y_t - F_t^s$: error in period t of the forecast made in period s

$\alpha = t - s$: forecast forward

$\beta = T - s$: forecast age

$\pi(\alpha)$: error power according to the forecast forward

$\mu(\alpha)$: multiplicative error factor according to the forecast forward

$\lambda(\beta)$: multiplicative error factor according to the forecast age

The RHWE (Rolling Horizon Weighted Error) criterion for the forecasting models selection is defined as follows:

$$RHWE = \sum_{t} \sum_{s} \left| e_t^s \right|^{\pi(\alpha)} \cdot \mu(\alpha) \cdot \lambda(\beta) \tag{1}$$

where:

$$\pi(\alpha) \geq 1 \tag{2}$$

$$\sum_{\alpha} \mu(\alpha) = 1 \tag{3}$$

$$\sum_{\beta} \lambda(\beta) = 1 \tag{4}$$

The model which presents the smallest criterion value is chosen.

The proposed criterion is an extension of the classic selection criteria based on the minimum one-step out-of-sample error. Specifically, the RHWE is summarised as:

- MAE when $\alpha = 1$, $\pi(\alpha) = 1$, $\mu(\alpha) = 1$ and vector $\lambda(\beta)$ is homogeneous.
- MSE when $\alpha = 1$, $\pi(\alpha) = 2$, $\mu(\alpha) = 1$ and vector $\lambda(\beta)$ is homogeneous.

In Table 1, we may observe a numerical example of the RHWE calculation for the exponential smoothing model. Forecasting begins by calculating from period 11 and a 4-period horizon is foreseen. The forecasts made from periods 11 to 20 have been recorded, and the errors made have been calculated when this instant in time is reached. An error power equal to 1 has been considered, that is, a multiplicative factor of the error according to the forecasting forward which is inversely proportional to the degree of advance, and a multiplicative factor of the error according to the age of the forecast which is inversely proportional to the age.

Table 1
Example of the RHWE calculation for the exponential smoothing model

α	**0,4**	s=10		s=11		s=12		s=13		s=14		s=15		s=16		s=17		s=18		s=19	
t	Y_t	F_t	e_t	F_t	e_t	F_t	e_t	F_t	e_t	F_t	e_t	F_t	e_t	F_t	e_t	F_t	e_t	F_t	e_t	F_t	e_t
1	91	*91,0*		*91,0*		*91,0*		*91,0*		*91,0*		*91,0*		*91,0*		*91,0*		*91,0*		*91,0*	
2	99	91,0		91,0		91,0		91,0		91,0		91,0		91,0		91,0		91,0		91,0	
3	56	94,2		94,2		94,2		94,2		94,2		94,2		94,2		94,2		94,2		94,2	
4	89	78,9		78,9		78,9		78,9		78,9		78,9		78,9		78,9		78,9		78,9	
5	49	83,0		83,0		83,0		83,0		83,0		83,0		83,0		83,0		83,0		83,0	
6	63	69,4		69,4		69,4		69,4		69,4		69,4		69,4		69,4		69,4		69,4	
7	40	66,8		66,8		66,8		66,8		66,8		66,8		66,8		66,8		66,8		66,8	
8	58	56,1		56,1		56,1		56,1		56,1		56,1		56,1		56,1		56,1		56,1	
9	87	56,9		56,9		56,9		56,9		56,9		56,9		56,9		56,9		56,9		56,9	
10	56	68,9		68,9		68,9		68,9		68,9		68,9		68,9		68,9		68,9		68,9	
11	40	**63,7**	**23,7**	63,7		63,7		63,7		63,7		63,7		63,7		63,7		63,7		63,7	
12	92	**63,7**	**-28,3**	**54,2**	**-37,8**	54,2		54,2		54,2		54,2		54,2		54,2		54,2		54,2	
13	60	**63,7**	**3,7**	**54,2**	**-5,8**	**69,3**	**9,3**	69,3		69,3		69,3		69,3		69,3		69,3		69,3	
14	42	**63,7**	**21,7**	**54,2**	**12,2**	**69,3**	**27,3**	**65,6**	**23,6**	65,6		65,6		65,6		65,6		65,6		65,6	
15	54			**54,2**	**0,2**	**69,3**	**15,3**	**65,6**	**11,6**	**56,2**	**2,2**	56,2		56,2		56,2		56,2		56,2	
16	25					**69,3**	**44,3**	**65,6**	**40,6**	**56,2**	**31,2**	**55,3**	**30,3**	55,3		55,3		55,3		55,3	
17	20							**65,6**	**45,6**	**56,2**	**36,2**	**55,3**	**35,3**	**43,2**	**23,2**	43,2		43,2		43,2	
18	14									**56,2**	**42,2**	**55,3**	**41,3**	**43,2**	**29,2**	**33,9**	**19,9**	33,9		33,9	
19	57											**55,3**	**-1,7**	**43,2**	**-13,8**	**33,9**	**-23,1**	**25,9**	**-31,1**	25,9	
20	20													**43,2**	**23,2**	**33,9**	**13,9**	**25,9**	**5,9**	**38,4**	**18,4**

	T-s	10	9	8	7	6	5	4	3	2	1	
	λ	0,02	0,04	0,05	0,07	0,09	0,11	0,13	0,15	0,16	0,18	
μ	0,4	0,17	0,55	0,2	0,69	0,08	1,32	1,18	1,16	2,03	1,34	
	0,3	0,15	0,06	0,45	0,25	0,85	1,16	1,11	1,01	0,29		
	0,2	0,01	0,09	0,17	0,59	0,66	0,9	0,35	0,4			
	0,1	0,04	0	0,24	0,33	0,38	0,02	0,3				**18,54**

Once the real values of the 9 forecasting periods are known for the described example, the RHWE value is 18,54 for a smoothing factor of 0,4. If the criterion for the selection of the smoothing factor were applied (the set of models considered would be of an exponential smoothing kind with varying smoothing factors), the criterion selected would be that which would provide a lower value. For this specific case, the smoothing factor which minimises the criterion is 0,56 whose value is 18,06.

6. Experiments

For the purpose of testing the proposed selection criterion, a set of forecasting models of various kinds has been chosen which range from the more classical models to those that use mathematical programming to establish their parameters.

There are two ways to establish the proposed selection criterion goodness: applying the proposed selection criterion goodness to other selection criteria, or applying it to a set of known time series with the out-of-sample errors from different forecasting models that have been documented. Even though the criterion has been used to check other previously tested selection criteria in the real case, the monthly M3 Competition time series was opted to perform the experiments whose objective was to verify the results obtained by the criterion proposed in relation to the results obtained from the models which participated in the competition.

6.1 Set of time series forecasting models

Different time series forecasting models have appeared in recent decades which may be

classified into various families. Among these we find the family of the exponential smoothing models (Montgomery *et al.*, 1990), which have been widely used in the practice, along with the ARIMA models (Box & Jenkins, 1970), which have been extensively used in the research. Despite the wide range of models, experience has shown that there is no forecasting model which works better than the rest in any given situation (Collopy & Armstrong, 1992). Therefore, a suitable selection of the model to be used for each time series may increase forecasting accuracy.

Depending on the characteristics of a specific time series, some forecasting models will provide better results than others. But, since the aim of this study is to analyse how a criterion to select time series forecasting models works, the definition of this set should not affect the results. For this reason, only two different forecasting models have been chosen to experiment with the M3 Competition monthly time series:

- Classical multiplicative decomposition.
- Weighted simple moving average (using a quadratic programming model to determine the weights minimizing the in-sample errors).

In a real application (as described in section 5.3), the number of models to consider must be large, and these models must present complementary performances for the time series that they have to deal with. Furthermore, the set of models to be considered must cover the different performances (seasonality, trend, intermittence, randomness, etc.) that may be noted in the time series to be dealt with.

6.2 Applying M3 Competition monthly time series

In order to validate the selection criterion of the time series forecasting models proposed, it was decided firstly to experiment with a data set on which the out-of-sample errors obtained with several forecasting models would have been calculated. To this end, the 1.428 monthly time series from the last international Makridakis competition (the M3 Competition) (Makridakis & Hibon, 2000) were selected. Other studies on the levels of accuracy of forecasting models may be consulted at Reid (1975), Newbold & Granger (1974), Makridakis & Hibon (1979), (Makridakis *et al.*, 1982), (Makridakis *et al.*, 1993).

The parameterisation implemented for this experiment was:

$$\alpha = 18 \tag{5}$$

$$\pi(\alpha) = 1 \tag{6}$$

In other words, an 18-period process in the forecasting (the forecast of the following 18 months in the M3 Competition were applied to the last time series value) and an error power equal to the unit (one of the most significant error measurements in the M3 was sMAPE).

In order to build the data for the model selection applying the RHWE criterion, an ex-ante simulation has been performed by assuming that the 18 final values of each time series were unknown, and by including a value of the time series in each calculation, which means a total of 19 forecasts for each time series. The 2 aforementioned forecasting models have been applied to the 1.428 time series of the M3 Competition, which gives a number of 2.856 forecasts, and 18 months of forecasting are calculated for each one. Upon obtaining

the out-of-sample errors on the time series values, which are supposedly unknown, a total of 171 errors are calculated for each forecast. Then, up to 488.376 out-of-sample errors are obtained.

Because of the objective of the experiment was to reach a good performance in the M3 Competition, the multiplicative factor of the error according to the forecasting forward $\mu(\alpha)$ and the multiplicative factor of the error according to the age of the forecast $\lambda(\beta)$ where estimated by means a mathematical programming model minimizing the average sMAPE of the final 18 values, performing a similar experiment as detailed in the previous paragraph, but considering the 18 final values of each time series as out-of-sample.

Table 2 presents a comparison of the accuracy of the forecasting models which participate in the M3 Competition, in which the RHWE criterion has been included at the end of the table. This is in fact the error measure according to the sMAPE not only for the different forecasting horizon points, and also for different averages. It is seen that the RHWE criterion overcomes the best models in the competition, and its good performance particularly stands out in the averages of the forecasting horizons. The weighted simple moving average was selected for the 70% of the time series and the classical multiplicative decomposition for the 30%.

Table 2
sMAPE of the RHWE criterion compared with the results of the M3 Competition. *Adapted from (Makridakis & Hibon, 2000)*

	Forecasting Horizons										Average of Forecasting Horizons					
METHODS	1	2	3	4	5	6	8	12	15	18	1-4	1-6	1-8	1-12	1-15	1-18
NAIVE2	15,00	13,50	15,70	17,00	14,90	14,70	15,60	16,00	19,30	20,70	15,30	15,13	15,29	15,57	16,18	16,91
SINGLE	13,00	12,10	14,00	15,10	13,50	13,10	13,80	14,50	18,30	19,40	13,55	13,47	13,60	13,83	14,51	15,32
HOLT	12,20	11,60	13,40	14,60	13,60	13,30	13,70	14,80	18,80	20,20	12,95	13,12	13,33	13,77	14,51	15,36
DAMPEN	11,90	11,40	13,00	14,20	12,90	12,60	13,00	13,90	17,50	18,90	12,63	12,67	12,85	13,10	13,77	14,59
WINTER	12,50	11,70	13,70	14,70	13,60	13,40	14,10	14,60	18,90	20,20	13,15	13,27	13,52	13,88	14,62	15,44
COMB S-H-D	12,30	11,50	13,20	14,30	12,90	12,50	13,00	13,60	17,30	18,30	12,83	12,78	12,92	13,11	13,75	14,48
B-J automatic	12,30	11,70	12,80	14,30	12,70	12,60	13,00	14,10	17,80	19,30	12,78	12,73	12,89	13,21	13,96	14,81
AUTOBOX-1	13,00	12,20	13,00	14,80	14,10	13,40	14,30	15,40	19,10	20,40	13,25	13,42	13,71	14,10	14,93	15,83
AUTOBOX-2	13,10	12,10	13,50	15,30	13,30	13,80	13,90	15,20	18,20	19,90	13,50	13,52	13,76	14,16	14,86	15,69
AUTOBOX-3	12,30	12,30	13,00	14,40	14,60	14,20	14,80	16,10	19,20	21,20	13,00	13,47	13,89	14,43	15,20	16,18
ROBUST-TREND	15,30	13,80	15,50	17,00	15,30	15,60	17,40	17,50	22,20	24,30	15,40	15,42	15,89	16,58	17,47	18,40
ARARMA	13,10	12,40	13,40	14,90	13,70	14,20	15,00	15,20	18,50	20,30	13,45	13,62	14,00	14,41	15,08	15,84
AutomatANN	11,60	11,60	12,00	14,10	12,20	13,90	13,80	14,60	17,30	19,60	12,33	12,57	12,92	13,42	14,13	14,93
FLORES-PEARC1	12,40	12,30	14,20	16,10	14,60	14,00	14,60	14,40	19,10	20,80	13,75	13,93	14,22	14,29	15,02	15,96
FLORES-PEARC2	12,60	12,10	13,70	14,70	13,20	12,90	13,40	14,40	18,20	19,90	13,28	13,20	13,33	13,53	14,31	15,17
PP-Autocast	12,70	11,70	13,30	14,30	13,20	13,40	14,00	14,30	17,70	19,60	13,00	13,10	13,37	13,72	14,36	15,15
ForecastPRO	11,50	10,70	11,70	12,90	11,80	12,30	12,60	13,20	16,40	18,30	11,70	11,82	12,06	12,46	13,09	13,86
SMARTFCS	11,60	11,20	12,20	13,60	13,10	13,70	13,50	14,90	18,00	19,40	12,15	12,57	12,90	13,51	14,22	15,03
THETAsm	12,90	12,20	13,60	14,30	14,10	14,30	14,00	14,20	17,60	19,10	13,25	13,57	13,85	14,06	14,56	15,26
THETA	11,20	10,70	11,80	12,40	12,20	12,40	12,70	13,20	16,20	18,20	11,53	11,78	12,13	12,50	13,11	13,85
RBF	13,70	12,30	13,70	14,30	12,30	12,80	13,50	14,10	17,30	17,80	13,50	13,18	13,40	13,67	14,21	14,77
ForcX	11,60	11,20	12,60	14,00	12,40	12,20	12,80	13,90	17,80	18,70	12,35	12,33	12,46	12,83	13,60	14,45
AAM1	12,00	12,30	12,70	14,10	14,00	14,00	14,30	14,90	18,00	20,40	12,78	13,18	13,63	14,05	14,78	15,69
AAM2	12,30	12,40	12,90	14,40	14,30	14,20	14,50	15,10	18,40	20,70	13,00	13,42	13,87	14,25	15,01	15,93
RHWE	11,46	10,48	11,77	12,68	11,14	12,33	12,09	12,85	15,61	16,51	11,60	11,64	11,80	12,09	12,67	13,31

Figure 1 shows the average value of the forecast age and forecast forward parameters for all the series. The one-step more recent error obtains the greatest (in average) weight for both parameters. But also wide horizons and past forecasting are taken into account with weights of 40% the greatest one. It seems that more recent forecasts are more important than short horizon forecasts and vice versa for older forecasts and wide horizon forecasts.

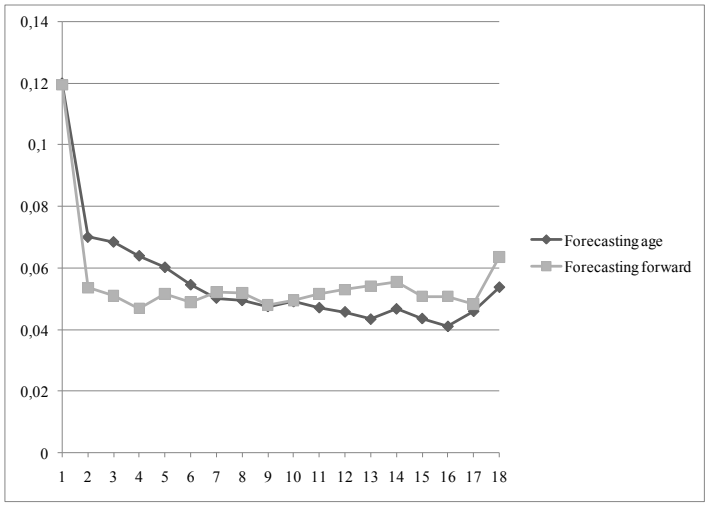

Fig. 1. Forecast age and forecast forward averages.

In order to learn the effect of the selection criterion, a similar experiment was done where the one-step MAE was used as the selection criterion. This case was placed in 17th position, which demonstrates that the out-of-samples errors made in wide horizons and obtained at different times greatly improve the performance of the selection criterion.

6.3 Application to a real case

The RHWE selection criterion of time series forecasting models has been applied to a real case in a firm that produces steel for construction. The RHWE criterion forms part of the expert system of a demand forecasting tool which automatically calculates demand forecasts on tens of thousands of time series of a very wide range (according to its characteristics, such as its seasonality, trend, intermittence, randomness, length, changes in pattern, outliers, etc.).

The expert system can chose for a set of 24 different forecasting models of various kinds which range from the more classical models to those that use mathematical programming to establish their parameters:

- Decomposition models: classical additive or multiplicative decomposition; Theta Model (Assimakopoulos & Nikolopoulos, 2000)
- Exponential Smoothing Models: Simple smoothing; Holt (Holt *et al.*, 1960); Holt-Winter (additive or multiplicative) (Winters, 1960); Croston (Croston, 1972); Syntetos-Boylan (Syntetos & Boylan, 2005)
- Moving Average Models: Simple moving average; Weighted simple moving average
- ARIMA Models (Box & Jenkins, 1970): Tramo/Seats (Gomez & Maravall, 2001)

The "classical" calculation models have been used in these models. For those which require the estimation of one parameter, or of several, mathematical programming models have been used (determinist and fuzzy) which minimise within-sample errors or out-of-sample errors. Furthermore, a combined model was included which applies ARIMA to the residual that has not been accounted for by the classical multiplicative decomposition method. A wide range of times series with different characteristics can be suitably

processed with this set of models.

Criterion parameterisation differs to that used for the experiment with the M3 Competition series since it is defined taking into account the relevance of the forecast accuracy along the horizon (12 months), depending the product family and the kind of decisions taken from this information. The result has been highly satisfactory as the RHWE criterion has, in general, coincided with the selection that an expert would have made.

7. Conclusions

This paper proposes a selection criterion for time series forecasting models by considering the out-of-sample errors recorded over time and for a specific horizon, whose importance is rated according to the distance from the instant in which the forecast is made and in accordance with the age of the forecast. This proposal also allows to define the degree of error power in accordance with the closeness of the period considered.

To validate the proposed method, two well known time series forecasting models has been used, which have been tested against the monthly times series of the M3 Competition, and which have overcome the models of the competition.

RHWE has been included in an expert system of a demand forecast tool, and it has been applied to a real case in a firm that produces steel for construction and which automatically forecasts on tens of thousands of time series on a monthly basis. The result has been highly satisfactory as the level of forecasting accuracy has increased in relation to the use of other selection criteria (for example, the lowest MSE).

Several future lines of research have been defined from the results obtained: a) to carry out experiments with different error powers, multiplicative factor vectors of the error according to the forecasting forward, and with multiplicative factor vectors of the error according to forecasting age; b) analysing experiments with variations in the parameters to determine the impact of each parameter; c) relation of the criterion parameters with the characterisations of the time series to be dealt with.

REFERENCES

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov, & F. Csaki (Eds.), *Second international symposium on information theory* (pp. 267–281). Budapest7 Akademiai Kiado.

Ali, Ö. G., Sayın, S., van Woensel, T., & Fransoo, J. (2009). SKU demand forecasting in the presence of promotions. *Expert Systems with Applications, 36*(10), 12340-12348

Arinze, B. (1994). Selecting appropriate forecasting models using rule induction, *Omega-International Journal of Management Science* 22 (6), pp. 647–658.

Armstrong, J S. (2001). Potential diffusion of expert systems in forecasting. *Technological forecasting and social change*, 67(1), 93.

Armstrong, J. S., Adya, M. & Collopy, F. (2001). Rule-based forecasting: Using judgment in time series extrapolation, in J. S. Armstrong (ed.), *Principles of Forecasting*. Norwell, MA: Kluwer Academic Press.

Armstrong, J.S. & Collopy, F. (1992). Error measures for generalizing about forecasting methods:

Empirical comparisons, *International Journal of Forecasting*, Vol. 8, Issue 1, pp. 69-80

Assimakopoulos, V. & Nikolopoulos, K., (2000). The theta model: a decomposition approach to forecasting, *International Journal of Forecasting*, Vol.16, No. 4, pp.521–530.

Behnamian, J., & Fatemi Ghomi, S. M. T. (2010). Development of a PSO–SA hybrid metaheuristic for a new comprehensive regression model to time-series forecasting. *Expert Systems with Applications, 37*(2), 974-984

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

Brown, R.G. (1959). *Statistical Forecasting for Inventory Control*, McGraw-Hill, New York.

Ching-Chin, C., Ka Ieng, A. I., Ling-Ling, W., & Ling-Chieh, K. (2010). Designing a decision-support system for new product sales forecasting. *Expert Systems with Applications, 37*(2), 1654-1665

Coccari, R.L. & Galucci, C. (1984). Average two best forecasts can reduce forecasting risk. *Journal Business Forecasting*, Fall.

Collopy, F. & Armstrong, J.S. (1989). Toward computer-aided forecasting systems, in G. R. Widemeyer (Ed.), *DSS 89 Transactions*, TIMS College on Information Systems. Providence, RI, 103-119.

Collopy, F. & Armstrong, J.S. (1992). Rule-based forecasting: development and validation of an expert systems approach to combining time series extrapolations, *Management Science*, Vol. 38, No. 10, pp.1394–1414

Croston, J.D. (1972). Forecasting and stock control for intermittent demands, *Operational Research Quarterly*, Vol. 23, No. 3, pp. 289–303.

De Gooijer, J.G. & Hyndman, R.J.(2005). 25 Years of IIF Time Series Forecasting: A Selective Review. Tinbergen Institute Discussion Paper.

Efendigil, T., Önüt, S., & Kahraman, C. (2009). A decision support system for demand forecasting with artificial neural networks and neuro-fuzzy models: A comparative analysis. *Expert Systems with Applications, 36*(3, Part 2), 6697-6707

Eom, H. B., & Lee, S. M. (1990). A survey of decision support system applications (1971 april 1988). *Interfaces*, 20(3), 65-79.

Eom, S. B., Lee, S. M., Kim, E. B., & Somarajan, C. (1998). A survey of decision support system applications (1988-1994). *Journal of the Operational Research Society*, 49(2), 109-120.

Eom, S., & Kim, E. (2006). A survey of decision support system applications (1995-2001). *Journal of the Operational Research Society*, 57(11), 1264-1278.

Fildes, R. & Makridakis, S. (1995). The impact of empirical accuracy studies on time series analysis and forecasting, *International Statistical Review,* Vol. 63, No. 3, 289-308.

Fildes, R. (1989). Evaluation of aggregate versus individual forecast method selection rules, *Management Science* 35, pp. 1056-1065.

Flores, B.E & Pearce, S.L. (2000). The use of an expert system in the M3 competition, *International Journal of Forecasting* 16 pp. 485–496

Franses, P.H. & Koehler, A.B. (1998). A model selection strategy for time series with increasing seasonal variation, *International Journal of Forecasting* 14 (1998) 405–414

Gomez, V. & Maravall, A. (2001). Seasonal adjustment and signal extraction in economic time series, Chapter 8 in *A course in time series analysis*, ed. D. Peña, G.C. Tiao and R.S. Tsay, John

Wiley & Sons: New York.

Holt, C. C., Modigliani, F., Muth, J. F. & Simon, H. A. (1960). *Planning Production Inventories and Work Force*, Prentice-Hall, Englewood Cliffs, NJ.

Inoue, A. & Kilian, L. (2006). On the selection of forecasting models, *Journal of Econometrics* 130 273–306

Khashei, M., & Bijari, M. (2010). An artificial neural network (p, d, q) model for timeseries forecasting. *Expert Systems with Applications, 37*(1), 479-489

Kuo, R. J., & Xue, K. C. (1998). A decision support system for sales forecasting through fuzzy neural networks with asymmetric fuzzy weights. *Decision Support Systems*, 24(2), 105-126.

Liao, S H. (2005). Expert system methodologies and applications--a decade review from 1995 to 2004. *Expert systems with applications*, 28(1), 93.

Lin, C., & Lee, I. (2009). Artificial intelligence diagnosis algorithm for expanding a precision expert forecasting system. *Expert Systems with Applications, 36*(4), 8385-8390

Liu, H. (2009). An integrated fuzzy time series forecasting system. *Expert Systems with Applications, 36*(6), 10045-10053

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

Makridakis, S. & Hibon, M. (2000), The M3-Competition: results, conclusions and implications *International Journal of Forecasting*, 16, pp. 451–476.

Makridakis, S. (1990). Sliding simulation: a new approach to time series forecasting, *Management Science* 36, pp. 505-512.

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, E. &Winkler, R. (1982). The accuracy of extrapolation (time series) methods: results of a forecasting competition. *Journal of Forecasting* 1, 111–153.

Makridakis, S., Chatfield, C., Hibon, M., Lawrence, M., Mills, T., Ord, K., & Simmons, L. F. (1993). The M-2 Competition: a real-time judgmentally based forecasting study. *International Journal of Forecasting* 9, 5–23.

Makridakis, S., Wheelright, S.C. & Hyndman, (1998). *Forecasting: Methods and Applications*. Wiley.

Miller, A. (1990). *Subset Selection in Regression*. Chapman and Hall.

Montgomery, D.C., Johnson, L.A. & Gardiner, J.S. (1990). *Forecasting and Time Series Analysis*, McGraw-Hill, New York.

Newbold, P. & Granger, C.W.J. (1974). Experience with forecasting univariate time series and the combination of forecasts. *Journal of Royal Statistical Society* A 137, 131–165.

Nikolopoulos, K. & Assimakopoulos, V. (2003). Theta intelligent forecasting information system, *Industrial Management & Data Systems* 103/9 pp. 711-726

Pai, P., Yang, S., & Chang, P. (2009). Forecasting output of integrated circuit industry by support vector regression models with marriage honey-bees optimization algorithms. *Expert Systems with Applications,* 36(7), 10746-10751

Petrovic, D., Xie, Y., & Burnham, K. (2006). Fuzzy decision support system for demand forecasting with a learning mechanism. *Fuzzy Sets and Systems*, 157(12), 1713-1725

Prudencio R.B.C., Ludermir T.B. & Carvalho F.A.T. (2004). A Modal Symbolic Classifier for selecting time series models, *Pattern Recognition Letters* 25, pp. 911–921

Reid, D.J. (1975). A review of short term projection techniques. In: Gordon, H.D., Editor, 1975. Practical aspects of forecasting, *Operational Research Society*, London, pp. 8–25

Sayed, H. E., Gabbar, H. A., & Miyazaki, S. (2009). A hybrid statistical genetic-based demand forecasting expert system. *Expert Systems with Applications, 36*(9), 11662-11670

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6, 461– 464.

Shu-Hsien Liao. (2005). Expert system methodologies and applications—a decade review from 1995 to 2004. *Expert Systems with Applications, 28*(1), 93-103

Simkins, S. (1995). Forecasting with vector autoregressive (VAR) models subject to business cycle restrictions. *International Journal of Forecasting* 11, 569–583.

Singh, S. R. (2009). A computational method of forecasting based on high-order fuzzy time series. *Expert Systems with Applications, 36*(7), 10551-10559

Spedding, T.A. & Chan, K.K. (2000). Forecasting demand and inventory management using bayesian time series, *Integrated Manufacturing Systems*, Vol. 11, No. 5, pp. 331-339

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society*, B36, 111–147.

Syntetos, A.A. & Boylan, J.E. (2005). The accuracy of intermittent demand estimates, *International Journal of Forecasting*, 21, pp. 303–314.

Tashman, L.J. (2000). Out-of-sample tests of forecasting accuracy: an analysis and review, *International Journal of Forecasting*, Volume 16, Issue 4, Pages 437-450.

Weiss, A. A. & Anderson, A. P. (1984). Estimating time series models using relevant forecast evaluation criteria, *Journal of the Royal Statistical Society* 147,pp.484-487.

Winters, P.R. (1960). Forecasting sales by exponentially weighted moving averages, *Management Science*, 6, pp. 324–342.

Wong, Chong (1993) Utilization and Benefits of Expert Systems in Manufacturing A Study of Large American Industrial Corporations. *International Journal of Operations & Production Management*, Vol. 14 No. 1, 1994, pp. 38-49.