

A Comparison of Different Model Selection Criteria for Forecasting EURO/USD Exchange Rates by Feed Forward Neural Network

Cagatay Bal,^{1*} Serdar Demir,² and Cagdas Hakan Aladag³

Abstract—Feed Forward Neural Networks (FFNN) are widely used in time series forecasting and produces satisfactory results for various real world applications. Many researches have showed that FFNN models also provide their success on forecasting exchange rate time series. In this study, various FFNN models are utilized to forecast EURO/USD exchange rate time series and different model selection criteria are used to determine the best architecture for the data. For this purpose, EURO/USD time series observed weekly from January of 1999 to January of 2016 is forecasted by using FFNN. To select different architectures according to different performance criteria, a computer program was coded by using MATLAB. As a result of the implementation, all obtained results are presented and interpreted.

Keywords— Artificial Neural Networks, EURO/USD, Exchange Rates, Forecasting, Model Selection Criterion, Time Series.

I. INTRODUCTION

USING Artificial Neural Networks (ANN) as a computational model inspired by biological neurons become very popular for forecasting in recent years. Hu (1964) used the Widrow's adaptive linear network in weather forecasting. Lapedes and Farber (1987, 1988) reported successful results for forecasting by using ANN. Since its good performance for both linear and nonlinear time series forecasting, ANN appeals to researchers and practitioners.

Good models for ANN are the networks which perform well with optimal number of neurons for test data set. Error measurements (model selection criteria, performance criteria) deal with this issue via different aspects for selection the best model. There are many different error measurements using for model selection in literature. The most popular error measurements are Mean Square Error (MSE), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Akaike Information Criteria (AIC), Bayesian Information Criteria (BIC).

Qi and Zhang (2001) investigates the performances of some popular model selection criteria such as AIC, Modified AIC (AICC, AIC1, AIC2, AIC3, AIC4), BIC, Modified BIC (BIC1, BIC2, BIC3, BIC4), MAE, MAPE, MSE, RMSE.

Shcherbakov et al. (2013) examines commonly used

forecasting error measurements such as RMSE, MAPE, Mean Relative Absolute Error (MRAE), Mean Absolute Scaled Error (MASE) and many more measurements with their different variations.

Here, we evaluate the performances of the almost all model selection criteria which are considered in literature for forecasting EURO/USD exchange rates.

In next section, we give the fundamentals of ANN and exchange rates. Section 3 reviews the model selection criteria. Section 4 reports the information and results of a simulation study which is performed to evaluate the performances of the criteria. Finally, we give conclusions and some discussions in Section 5.

II. ANN FOR TIME SERIES FORECASTING

Economical time series (e.g., exchange rates, stock exchange prices, earthquake magnitudes) generally consist of complex structures which include both linear and non-linear data. Massive variability and uncertainty are often seen for the EURO/USD exchange rates. So it becomes much more important for successful forecasting and to minimize measurement errors for economic actors all around the world.

Deficiencies of classical methods make this process even harder. However, ANN has the learning ability for both linear and nonlinear time series. So this makes it a good alternative for forecasting exchange rates.

Qi & Zhang (2001) states that the comparative advantage of neural networks over more classical methods lies in the fact that they can flexibly model complex, possibly nonlinear relationship without any prior assumptions about the underlying data-generation process.

ANN comprises algorithms which mimic the mechanism of human brain. This mechanism is producing new knowledge by learning the mysterious relations in data. The components of ANN are network architecture, learning algorithm, and activation function. Selecting the types of these components is too important for performance of ANN.

Since there isn't general rules for choosing the best architecture in the literature, many different architectures are examined to get appropriate results in this study. For input and hidden layers neuron numbers varies from 1 to 12 respectively so it becomes 144 different structures. It has no affect the total number of different structures to have only 1 neuron for output layer. Figure 1 shows FFNN architecture.

^{1,2} Department of Statistics, Mugla Sitki Kocman University, Mugla, Turkey

³ Department of Statistics, Hacettepe University, Ankara, Turkey.

*Corresponding author. Tel.: +905398973233; email: cagataybal@mu.edu.tr

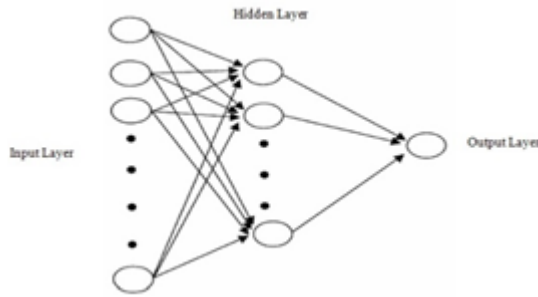


Fig. 1: FFNN architecture

Learning ability from a source of knowledge is one of the most important characteristic of ANN. Learning process is other crucial component of ANN and it means to obtain best weights between given inputs. Back Propagation learning algorithm first was introduced by Werbos (1974) and after the capabilities of ANN reveals, many studies have been made and many more on the process of working.

Another crucial component of ANN is activation function. Activation function provides the non-linear mapping between input and output. The performance of networks depends on the appropriate choice of activation function. In general, the activation function represents the non-linearity degree which is critical in ANN applications. Moreover, every neuron in the network has an input and output. The output of neuron contains weighted sum of inputs. But for the input layer, neurons don't have any weight. Because this layer is just for receive inputs from given data and delivers it into network. Before producing the output of the neuron, weighted sum of inputs which names *net* value, fits in a function than produces output delivers to another neuron as an input. We used *hyperbolic tangent* function for hidden layer and *linear* function for the output layer as an activation function. These functions are the most commonly used in the literature. Figure 2(a)&(b) shows both functions.



Fig. 2. Activation Functions: (a) Linear function; (b) Hyperbolic tangent function.

In this study, generated networks consist of three layers as input, hidden and output layers. Number of neurons for input and hidden layers varies from one to twelve but for the output layer only one neuron is used. For its fast processing and wide usage in literature, Levenberg-Marquardt Back Propagation (LMBP) is used for learning algorithm.

III. MODEL SELECTION CRITERIA

Determining the best architecture of ANN is too important issue in the applications (Egrioglu, Aladag & Gunay, 2008). Every performance measure assess forecasting error from different aspects and various model selection criteria have

been used for the determining the best architecture of ANN.

We discuss all criteria in six subsections as absolute, percentage, symmetric, relative, scaled and others due to high number of measurements.

3.1 Absolute Errors

Absolute errors are defined as follows:

$$e_t = (y_t - \hat{y}_t) \tag{1}$$

Here, y_t is measured value and \hat{y}_t is predicted value at time t . In all formulas, n is used as sample size. The formulas of model selection criteria based on absolute errors are given in Table 1.

TABLE I
MODEL SELECTION CRITERIA BASED ON ABSOLUTE ERRORS.

Name of Criteria	Formula
Mean Absolute Error	$MAE = \text{mean}_{i=1,n} e_i $
Median Absolute Error	$MdAE = \text{median}_{i=1,n} e_i $
Geometric Mean Absolute Error	$GMAE = \text{gmean}_{i=1,n} e_i $
Mean Square Error	$MSE = \text{mean}_{i=1,n} (e_i^2)$
Root Mean Square Error	$RMSE = \sqrt{\text{mean}_{i=1,n} (e_i^2)}$
Fourth Root Mean Quadrupled Error	$R4MS4E = \sqrt[4]{\text{mean}_{i=1,n} (e_i^4)}$

The most important disadvantage of absolute error is dependency of scale. Different scaled object for the same forecasting task may lead to confusion and making the results inappropriate. Also this situation could make the comparison between studies or methods impossible. To overcome this, researches should scale every data before modelling.

3.2. Percentage Errors

Percentage errors are defined as follows:

$$p_t = (y_t - \hat{y}_t) / y_t \tag{2}$$

The formulas of model selection criteria based on percentage errors are given in Table 2.

TABLE II
MODEL SELECTION CRITERIA BASED ON PERCENTAGE ERRORS.

Name of Criteria	Formula
Mean Absolute Percentage Error	$MAPE = \text{mean}_{i=1,n} p_i $

Median Absolute Percentage Error	$MdAPE = \text{median}_{i=1,n} p_i $
Root Mean Square Percentage Error	$RMSPE = \sqrt{\text{mean}_{i=1,n} (p_i^2)}$
Root Median Square Percentage Error	$RMdSPE = \sqrt{\text{median}_{i=1,n} (p_i^2)}$

The most important disadvantage of percentage error is division by zero when measured value is equal to zero. Some modification should be added to measures to overcome this situation. Another important disadvantage is non-symmetry of error which means values may change according to predicted value is bigger or smaller than measured ones. Therefore results become biased.

3.3 Symmetric Errors

Symmetric errors are defined as follows:

$$s_t = \frac{|y_t - \hat{y}_t|}{(y_t + \hat{y}_t)} \quad (3)$$

The formulas of model selection criteria based on symmetric errors are given in Table 3.

TABLE III

MODEL SELECTION CRITERIA BASED ON SYMMETRIC ERRORS	
Name of Criteria	Formula
Symmetric Mean Absolute Percentage Error	$SMAPE = \text{mean}_{i=1,n} (s_i)$
Symmetric Median Absolute Percentage Error	$SMdAPE = \text{median}_{i=1,n} (s_i)$

Although this measures names actually these measures are not symmetric (Koehler, 2001). Same as percentage errors disadvantage, division by zero hardly but may occur for symmetric errors.

3.4 Relative Errors

Relative errors are defined as follows:

$$r_t = \frac{|y_t - \hat{y}_t|}{(y_t - f_t^*)} \quad (4)$$

Here, $f_t^* = y_{t-l}$ can be called as reference model value or prediction. Common usage of reference model is lagged measured value at time t and l lag. In this study we used $f_t^* = y_{t-1}$ as reference model.

The formulas of model selection criteria based on relative errors are given in Table 4.

TABLE IV

MODEL SELECTION CRITERIA BASED ON RELATIVE ERRORS.	
Name of Criteria	Formula
Mean Absolute Error	$MAE = \text{mean}_{i=1,n} e_i $

Median Absolute Error	$MdAE = \text{median}_{i=1,n} e_i $
Geometric Mean Absolute Error	$GMAE = \text{gmean}_{i=1,n} e_i $
Mean Square Error	$MSE = \text{mean}_{i=1,n} (e_i^2)$
Root Mean Square Error	$RMSE = \sqrt{\text{mean}_{i=1,n} (e_i^2)}$
Fourth Root Mean Quadrupled Error	$R4MS4E = \sqrt[4]{\text{mean}_{i=1,n} (e_i^4)}$

The most important disadvantage of relative error is division by zero when measured value and reference models value are the same.

3.5 Scaled Errors

Scaled errors are defined as follows:

$$sc_t = \frac{|y_t - \hat{y}_t|}{\frac{1}{n-1} \sum_{t=2}^n |y_t - y_{t-1}|} \quad (5)$$

The formulas of model selection criteria based on scaled errors are given in Table 5.

TABLE V

MODEL SELECTION CRITERIA BASED ON SCALED ERRORS.	
Name of Criteria	Formula
Mean Absolute Scaled Error	$MASE = \text{mean}_{i=1,n} sc_i $
Root Mean Square Scaled Error	$RMSSE = \sqrt{\text{mean}_{i=1,n} (sc_i^2)}$

3.6 Other Errors

Other errors which are not include other category are defined as follows:

TABLE VI

MODEL SELECTION CRITERIA BASED ON SCALED ERRORS	
Name of Criteria	Formula
Akaike Information Criteria	$AIC = \log \left(\frac{\sum_{t=1}^n (y_t - \hat{y}_t)^2}{n} \right) + \frac{2 \cdot m}{n}$
Bayesian Information Criteria	$BIC = \log \left(\frac{\sum_{t=1}^n (y_t - \hat{y}_t)^2}{n} \right) + \frac{m \cdot \log(n)}{n}$
Nash Sutcliffe Efficiency	$NS = 1 - \frac{\sum_{t=1}^n (y_t - \hat{y}_t)^2}{\sum_{t=1}^n (y_t - \bar{y})^2}$

In Table 6, parameter m is equal to total number of weights in ANN. As seen above, each model selection criteria has advantages and disadvantages. It can be said that each measure based on *mean* has the same disadvantage as having high influences for outliers. In addition, it can be said that

each measure based on *median* has the same disadvantages as long time calculation if data set is large. AIC and BIC penalize the models with much parameters, so they may be over penalize because ANN consists of many parameter.

IV. APPLICATION

An application study is performed to evaluate the

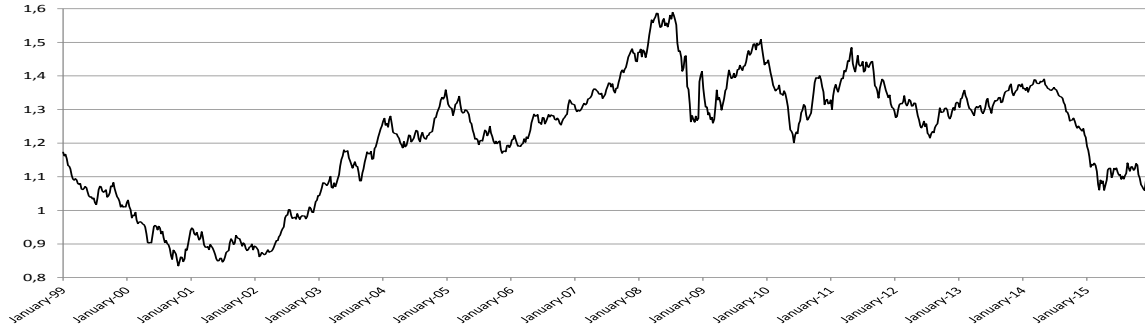


Fig.3. Graph of EURO/USD exchange rate between the January 1999 and January 2016.

Although neural networks have been successfully used for numerous forecasting applications, several issues in ANN model building still have not been solved (Zhang et al. 1998). Despite ANN can handle this type of data easily, the components must be built in optimized form.

In the ANN literature, there are two types of model selection strategies to ameliorate overfitting issue. The first and used method for this study is cross validation approach which is dividing data into parts. Generally data division can be done into three parts as training, validation and test. Training and validation sets are for ANN model development and test set is for measuring the performance. We use this strategy for implementation of our study without validation set. EURO/USD exchange rate time series data divided into two sets as training and testing. First 836 observations of data used for training and last 52 observations used for testing. The performance of ANN is evaluated by comparing each model selection criteria for the test set.

The second method is coming from classical time series approaches. One certain information criteria like AIC or BIC tries to find the best forecasting model within the training process. However the disadvantage of this method is ignoring the alternative models could be better than actual model selected by the information criteria. Also experiences show that models selected by training errors give poor forecasting performance. Oppositely models selected by test errors give well forecasting performance. These results explain the second method could lead the ANN overfitting and first method could lead the ANN learning.

In application progresses, much different architecture achieved to get appropriate results. For input and hidden layers neuron numbers varies from 1 to 12 respectively so it becomes 144 different structures for one circle. Due to having only 1 neuron for output layer it doesn't affect the total number of different structures. Application progress comprises 100 circles. For every circle, best model for each criterion selected among 144 architectures. Finally the 100

performances of the model selection criteria over forecasting the EURO/USD exchange rate. The end-of-day exchange rates were taken from official web site of *Republic of Turkey Central Bank*. The data are weekly average exchange rates between the January 1999 and January 2016 and consist of 888 observations. Figure 3 shows the graph of the data.

best models selected by each criterion have been ready for comparison.

TABLE VI
STATISTICS OF 1 STEP AHEAD FORECASTS FOR EACH CRITERION

Measure	N	Range	Minimum	Maximum	Mean	Std. Deviation
MAE	100	7,0202	-5,7794	1,2408	0,8738	0,9168
<u>MdAE</u>	100	1,1314	0,4964	1,6278	1,0188	0,1168
GMAE	100	36,5008	-23,1330	13,3678	0,8549	2,7391
MSE	100	4,7578	-3,3643	1,3935	0,9984	0,4476
RMSE	100	4,7578	-3,3643	1,3935	0,9984	0,4476
R4MS4E	100	9,0176	-3,3643	5,6533	1,0324	0,6479
MAPE	100	7,1729	-5,7794	1,3935	0,9070	0,8540
<u>MdAPE</u>	100	3,3978	-1,7700	1,6278	1,0050	0,3031
RMSPE	100	6,6781	-3,3643	3,3137	1,0153	0,5035
<u>RMdSPE</u>	100	3,3978	-1,7700	1,6278	1,0050	0,3031
SMAPE	100	7,1729	-5,7794	1,3935	0,9094	0,8384
<u>SMdAPE</u>	100	3,3978	-1,7700	1,6278	1,0012	0,3111
MRAE	100	7,0202	-5,7794	1,2408	0,8760	0,9174
<u>MdRAE</u>	100	1,1314	0,4964	1,6278	1,0268	0,1209
GMRAE	100	36,5008	-23,1330	13,3678	0,8549	2,7391
MASE	100	11,0259	-2,2145	8,8114	1,1115	0,9529
RMSSE	100	5,0528	0,6005	5,6533	1,0999	0,4675
NS	100	4,7578	-3,3643	1,3935	0,9984	0,4476
AIC	100	0,0044	1,0797	1,0841	1,0805	0,0016
BIC	100	0,0044	1,0797	1,0841	1,0805	0,0016

Statistics (range, minimum, maximum, mean, standard deviation) of 1 step ahead forecasts for each criterion over all 100 circles are given in Table 6. For each criterion, the 100 best models and 1 step ahead forecasts are obtained. Range and standard deviation specifies variability of the criteria. Hence, the criterion with minimum range and standard deviation of forecasts means better consistent criteria for model selection. It can be seen that every measure with *median* from each category has the minimum value of range and standard deviation. Oppositely, every measure with *geometric mean* from each category has the maximum value of range and standard deviation.

TABLE VII. THE BEST OF THE 100 BEST MODELS FOR EACH MEASURE.

Measure	Architectures	Error	Measure	Architectures	Error
MAE	5-9-1	0,0158	SMAPE	5-9-1	0,0431
MdAE	7-12-1	0,0098	SMdAPE	7-12-1	0,0267
GMAE	7-12-1	0,0076	MRAE	5-9-1	0,0159
MSE	5-9-1	0,0004	MdRAE	7-12-1	0,0098
RMSE	5-9-1	0,0199	GMRAE	7-12-1	0,0076
R4MS4E	7-11-1	0,0249	MASE	6-8-1	0,4919
MAPE	5-9-1	0,0433	RMSSE	5-10-1	0,6938
MdAPE	7-12-1	0,0269	NS	5-9-1	0,2855
RMSPE	5-9-1	0,0549	AIC	1-1-1	-7,5859
RmdSPE	7-12-1	0,0269	BIC	1-1-1	-7,5109

AIC and BIC as penalty based criteria can be seen as the best measure but because of their disadvantage of ignoring alternative architectures, only one model has been chosen for both criteria in each circle. So it is obvious to say that these criteria are not suitable to use for model selection in ANN.

The best of the best models/architectures and error values selected by each criterion within the 100 best models can be seen in Table 7.

V. CONCLUSION

It is important to decide appropriate criteria for choosing the best architecture. There are many performance criteria of model selection for forecasting time series in literature. All these criteria have both advantages and disadvantages such as consistency/inconsistency, sensitivity to outliers and extreme values, computing time.

In this study, the performances of model selection criteria are analyzed for forecasting EURO/USD exchange rate time series. The results shows that the model selection measures based on median of forecasts have lower variability than others whereas the measures based on geometric mean have higher variability. For these types of data with high volatility, it can be suggest to use the measures based on median (MdAE, MdAPE, SMdAPE, MdRAE) and not to use the measures based on penalizing (AIC, BIC).

REFERENCES

- [1] M.J.C., Hu, 1964, "Application of the adaline system to weather forecasting." Master Thesis, Technical Report 6775-1, Stanford Electronic Laboratories, Stanford, CA, June.
- [2] A., Lapedes, R., Farber, 1987, "Nonlinear signal processing using neural networks: prediction and system modeling." Technical Report LA-UR-87-2662, Los Alamos National Laboratory, Los Alamos, NM.
- [3] A., Lapedes, R., Farber, 1988, "How neural nets work." In: Anderson, D.Z., (Ed.), Neural Information Processing Systems, American Institute of Physics, New York, pp. 442-456.
- [4] M., Qi, G.P., Zhang, (2001), "An Investigation of Model Selection Criteria for Neural Network Time Series Forecasting", European Journal of Operational Research, 132 666-680.
- [5] M.V., Shcherbakov, et al. (2013), "A Survey of Forecast Error Measures." Information Technologies in Modern Industry, Education & Society, 171-176.
- [6] P.J., Werbos, 1974, "Beyond regression: new tools for prediction and analysis in the behavioral sciences." Ph.D. thesis, Harvard University.
- [7] E. Egrioglu, C.H. Aladag, S. Günay, "A new model selection strategy in artificial neural networks", Applied Mathematics and Computation, 195 (2008) 591-597.
- [8] A.B., Koehler, 2001, "The asymmetry of the sAPE measure and other comments on the M3-competition." International Journal of Forecasting, 17: 570-574.

- [9] G. Zhang, B.E. Patuwo, Y.M. Hu, "Forecasting with artificial neural networks: the state of the art", International Journal of Forecasting 14 (1998) 35-62.
- [10] P., Maca, P., Pech, J., Pavlasek (2014), "Comparing the Selected Transfer Functions and Local Optimization Methods for Neural Network Flood Runoff Forecast". Mathematical Problems in Engineering, Volume 2014, Article ID 782351, 10 pages.
- [11] C.H. Aladag, E. Egrioglu, S. Günay, M.A. Basaran, "Improving weighted information criterion by using optimization", Journal of Computational and Applied Mathematics 233 (2010) 26832687.