Stock index returns forecasting using a hybrid neural network model

K.J.H. Dekker
412487
Bachelor Thesis: Econometrics and Operations Research
Erasmus School of Economics, ERASMUS UNIVERSITY ROTTERDAM

Supervisor: S.L.H.C.G. Vermeulen

Second assessor: S.C. Barendse

5th July 2018

Abstract

Stock returns are likely to consist of linear parts as well as nonlinear parts over time. Autoregressive or moving average models may be suitable for capturing linearly behaving parts, whereas nonlinear smooth transitioning or artificial neural network models can be used to describe the nonlinear habits. In general, none of the aforementioned individual models is able to capture both linearity and nonlinearity of time series completely. To overcome this, my study combines two models from the linear and nonlinear field into a hybrid model to describe and forecast the returns on the Spanish Ibex-35 stock index. Recursive and non-recursive one-step-ahead and multi-step-ahead forecasting methods are utilised over three forecast horizons to measure and compare the performance of the hybrid model relative to the individual models. I find that the hybrid model is able to surpass the individual models for short horizon one-step-ahead forecasts. The hybrid model seems particularly useful for predicting the correct sign of the returns, which indicates that the combined model can be of interest to incorporate in trading strategies.

1 Introduction

Which model can forecast the future returns of a stock index as accurate as possible? That is a question that many investors would like to answer. Predicting the returns can be of interest when incorporating a trading strategy over a longer horizon. A model that can describe the movement of stock returns universally has yet to be found.

In the field of time series modelling and forecasting there are several different approaches. Most traditional statistical models include exponential smoothing, autoregressive (AR) and moving average (MA) models. These models are linear in a sense that the predictions of a variable in the future are constrained to be linearly dependent on the observations in the past. Because of the relative simple nature in understanding and implementation, the performance of linear models has been studied quite elaborately during the past decades.

One of the most important and extensively used linear time series model is the autoregressive integrated moving average (ARIMA) model. This model gained popularity due to its statistical properties combined with the model building process, which follows a well-known Box-Jenkins methodology [5]. Besides that, several different methods like the exponential smoothing model can be implemented by ARIMA models. One big advantage of ARIMA models is their flexibility. ARIMA models can capture pure AR, pure MA and combined AR and MA (ARMA) series. However, the greatest drawback of these models is their assumed linear form. Due to this assumption, no nonlinear patterns can be captured by the ARIMA models. Therefore the approximation of linear models to explain possibly nonlinear real-word observations are not always satisfactory.

In many studies the efficient market hypothesis is questioned [11, 17, 34]. The assumption that complex asset price movements can be described by a random walk model is debatable. As stated, time series do not always exhibit a linear relation and may very well be nonlinear. This is well-grounded because the linear models are not always capable of explaining several agency aspects. These aspects include the possibility that not all agents receive information simultaneously, the importance of differences in negotiation times or how those agents with more complex algorithms can make better-informed decisions. The conclusions of papers which investigate the efficient market hypothesis support both a theoretical as a practical interest in nonlinear financial time series models.

Broadly used nonlinear models include bilinear models, autoregressive conditional heteroscedastic (ARCH) models and smooth transition autoregressive (STAR) models [4, 9]. These models offset linear models when forecasting nonlinearly behaving time series. Although this improvement has been noticed, the gain of forecasting with nonlinear models over general linear models is limited [12]. Because of the fact that nonlinear models are structured for certain specific nonlinear patterns, they are not able to correctly capture multiple types of nonlinearities simultaneously in time series. This led to the more recent rise of artificial neural network (ANN) models, which have been suggested as a useful nonlinear alternative for forecasting time series. The strength of these ANN models lays within their flexible nonlinear modelling capacity. These models have the capability to capture multiple types of nonlinearities which may exhibit in time series data.

Both ANN and ARIMA models are successful models in the nonlinear and linear domain respectively. Yet, none of them is considered as a universal model which is suitable under all

circumstances. If one tries to estimate the parameters of an ARIMA model to account for a complex nonlinear problem, it is very likely that this model would not capture this nonlinearity properly as the data generating process differs. In addition, fitting an ANN model for clearly linear data may be totally inappropriate as well. Moreover, real-world time series data often consist of complex movements, both linear as well as nonlinear, which can not be captured well by one of the aforementioned models separately. This leads to a model selection problem with divergent solutions. Hence, a combination of these two models could yield a viable solution in the form of a model, which is able to handle both linearity and nonlinearity adequately.

Theoretical and empirical evidences in current literature [25, 27, 28, 35] suggest that, by making use of dissimilar models or models that capture disjoint characteristics, a hybrid model will give lower generalization variance or error. The aim of a combined model is that the risk of using an inappropriate model for the data is minimized or at least reduced. Combining thus reduces risk of an erroneous model and obtains results which are more in line with the underlying data and are on average more accurate.

This paper extends the work of Pérez-Rodríguez et al. [29] by making use of a hybrid ARIMA-ANN model to forecast Ibex-35 stock index returns. They find that ANN models have a better model fit over STAR models, as the former offers slightly higher trading profits in terms of one-step-ahead forecasts as in Sharpe ratio. However, they are unable to determine whether the ANN models are superior over other (non)linear models in the multi-step-ahead case. Empirical results support that making use of a hybrid model for financial market forecasting can effectively improve accuracy, such that the hybrid model can be used as an alternative to financial market forecasting tools. Especially when only a short-term time series is investigated, the hybrid model performs relatively well [37].

In this paper I will propose a method to combine the linear and nonlinear aspects into a hybrid model for better modelling and forecasting purposes. In section 2 I will review the current knowledge of forecasting with the individual models as well as with some hybrid models. Section 3 explains the model building process and the forecast methods. The proposed methodology of the hybrid model consists of two stages. In the first stage a linear ARIMA model is fitted in order to identify the existing linear components of the data. In the second stage I will introduce an existing nonlinear ANN model to account for nonlinearity in the sample data. This is followed in section 4 by a description of the data that I use for my analysis. In section 5 the performance of this hybrid model will be evaluated in comparison with the performance of the separate linear and nonlinear models in terms of statistical criteria such as goodness of forecast measures, proportion of times that the signs are correctly predicted, the directional accuracy test and tests for equality of accuracy of competing forecasts. I produce forecasts of the models using (recursive) one-step-ahead as well as multi-step-ahead methods. Section 6 and 7 respectively conclude and discuss my findings and give directions for further research.

2 Literature

Forecasting time series data is one the most prominent and widely discussed topics within the fields of statistical analysis. The possibility of exploiting linearities as well as nonlinearities within these datasets to improve forecasting performances over short and long horizons could be of great importance.

In many studies of empirical finance, the efficient market hypothesis is questioned [11, 17, 34].

According to those studies it is doubtful that the random walk model is a reasonable description of asset price movement and that linear modelling techniques are able to capture complex movements of asset prices [11, 34]. Another often proposed model for capturing financial time series data is the ARIMA model, which gained popularity due to its statistical properties as well as the Box-Jenkins building methodology. However, the performance of these linear models can be poor for seemingly nonlinear data, as the models lack the capability of capturing nonlinear patterns.

As a result of this lack, the interest in nonlinear models is fueled. Those models are able to capture more complex movements over time. Among the first were Lapedes and Farber [22] to attempt to model a nonlinear time series with artificial neural networks. Later on, De Groot and Wurtz [15] present a detailed analysis of forecasting a univariate time series by making use of a feedforward neural network for two nonlinear time series, the annual amount of sunspots and a deterministic chaotic time series. In addition, several large forecasting competitions suggest that the neural network can be a promising addition to the forecasting toolbox [2].

One of the major developments in neural network modelling and forecasting was the introduction of combined models. The basic idea behind this multi-model approach is that the ideal unique capability of the underlying models can capture different patterns in the data. Both theoretical as empirical findings suggest that the predictive performance of a hybrid model offsets the individual model, especially when the models worked with are quite different [3, 36].

Among the first incorporations of such a combined hybrid model is Zhang [37], who proposes a combination of the linear ARIMA model along with a feedforward ANN model. This model is then used empirically with three often used real-world nonlinear time series datasets, namely the sunspot data, Canadian lynx data and the British Pound/US Dollar exchange rate. His findings indicate that a combined model can be effective to improve forecasting accuracy achieved by the separate models.

Also Wang and Meng [33] find that the hybrid form of a combined ANN and ARIMA model improves forecasts of the energy consumption of a province in China relative to those obtained by the separate models. Furthermore, Khashei et al. [20] propose a novel hybrid model in order to overcome limitations of ANN models to yield even more accurate results by making use of a fuzzy regression model.

These positive findings for different non-financial nonlinearly behaving time series, enhanced interest in the application in the financial sector. Khashei et al. [19] presented a hybrid ARIMA-ANN approach to model financial market predictions. Also Tseng et al. [32] propose a hybrid model called the SARIMABP, which contained the seasonal ARIMA (SARIMA) model and a backpropagation ANN model to forecast total production value for the Taiwanese machinery industry.

Concluding, the combination of an ARIMA model with an ANN model seems to be very fruitful in multiple fields of research that make use of (possibly nonlinear) time series data. This research focusses on the forecast comparison with other individual (non)linear models. I also investigate that if the hybrid model provides superior results, to what extent this model can improve time series data forecasting in a financial data framework.

3 Methodology

In this section the basic concepts and modelling approaches of the different models are briefly reviewed. Models of interest include the linear ARIMA model, nonlinear STAR model, artificial neural network multilayer perceptron (MLP) model and a hybrid model combining ARIMA and MLP models. After establishing the models, forecast methods are used in order to assess statistical and economic criteria of out-of-sample forecasts. The dependent variable of interest is the return (r_t). Lagged observations and lagged error terms are the only explanatory variables being used in the models.

3.1 ARIMA model

There is a broad range of approaches to model time series. A traditional linear model that has been used extensively over the last 50 years is the ARIMA model. An ARIMA model consists of three parts, the AR function regresses on historic values of the index, the MA function regresses on a purely random process and the integrated (I) part ensures that the time series is stationary by differencing. Whereas other models are able to capture nonlinearity, the ARIMA model deals with non-stationary linear components.

The ARIMA model assumes that the future observation of a variable is a linear function of a couple of past observations and random errors. Stated otherwise, the underlying process which generates the time series data with mean μ is of the form

$$\phi(B)\nabla^d(r_t - \mu) = \theta(B)\varepsilon_t, \tag{1}$$

where r_t and ε_t resemble the return of the index and the random error at time t. The functions $\phi(B) = 1 - \sum_{i=1}^{p} \phi_i B^i$ and $\theta(B) = 1 - \sum_{j=1}^{q} \theta_j B^j$ are the polynomials in B of degree p and q respectively, where ϕ_i (for $i = 1, 2, ..., p$) and θ_j (for $j = 1, 2, ..., q$) are the model parameters. Here p and q are integers which are often referred to as the order of the AR part and the MA part included in the model. Furthermore, $\nabla = (1 - B)$, where B is the backward shift operator. In (1) d is also an integer which is often referred to as the order of differencing. The random errors ε_t are assumed to be i.i.d. with zero mean and a constant variance σ^2.

The methodology as described by Box-Jenkins [5] includes three iterative stages of model identification, parameter estimation and diagnostic checking. The main principle of model identification is that if a time series is generated according to an ARIMA process, it should have some properties of theoretical autocorrelation. By comparing the empirical autocorrelation with the theoretical values, it is often possible to identify one or multiple potential models for the given time series data. As proposed by Box and Jenkins the autocorrelation function and the partial autocorrelation function of the observed data are useful basic tools for order identification of the ARIMA model. Other information criteria like the AIC or the BIC can also be used for order selection.

In the identification step the data is required to be a stationary time series, since stationarity is necessary for building an ARIMA model used for forecasting. Stationary time series are characterized by constant mean and autocorrelation structure over time. When the observed data does present a trend or heteroscedasticity, transformation in the form of differencing has to be applied in order to remove the trend and stabilize the variance. Once this is established, estimation of the parameters in the model is straightforward. For estimating the parameters, the overall measure of errors is minimized, which can be achieved by making use of a nonlinear

optimization procedure. The last step is a diagnostic check of the adequacy of the model. This boils down to checking whether the model assumptions which are made about the errors, ε_t, are satisfied. Residual plots and diagnostic statistics can be used to examine the goodness of fit of the proposed model for the historical data. If the model is not able to explain the data well, a new model should be identified which will be followed by repeating the steps of parameter estimation and model verification. The structure of the residuals and the diagnostic information may help to suggest an alternative, better model. This three stage approach for model building is typically repeated a couple of times until a satisfactory model is selected. This final model can then be used for the forecasting purposes.

3.2 STAR models

STAR models are especially useful if the data exhibit multiple regimes with potentially different dynamic properties, but with a smooth transition between those regimes [14, 31]. A simple first-order STAR model with two regimes is given by

$$r_t = \phi_{10} + \sum_{i=1}^{p} \phi_{1i} r_{t-i} + \left[\phi_{20} + \sum_{i=1}^{p} \phi_{2i} r_{t-i}\right] F_{t,d}(s_t; \gamma, c) + \varepsilon_t. \tag{2}$$

Here r_t represents the stock index return at t, ϕ_{ji} (for $j = 1, 2; i = 0, 1, 2, ..., p$) are the unknown parameters that correspond to each regime j. The function $F_{t,d}(s_t; \gamma, c)$ is the transition function, which is assumed to be twice differentiable and in the range [0, 1]. The parameter γ is the transition rate or the smoothness parameter and c stands for the threshold value, which represents the change from one regime to another. The number of lags of the transition variable is given by d. The transition function allows for switching between regimes and introduces non-linearity into parameters of the model. The transition variable, s_t, is usually defined as a linear combination of lagged returns r_t: $s_t = \sum_{i=1}^{d} \alpha_i r_{t-i}$. Specification of the transition function is mostly done by a first-order logistic function (LSTAR) or a first-order exponential function (ESTAR). The logistic specification (3) and the exponential speficiation (4) of the transition function are given by

$$F_{t,d}(s_t; \gamma, c) = \{1 + exp[-\gamma(s_t - c)]\}^{-1}, \tag{3}$$

$$F_{t,d}(s_t; \gamma, c) = \{1 - exp[-\gamma(s_t - c)^2]\}. \tag{4}$$

For both specifications hold that $\gamma > 0$ and that the transition variable, s_t, can be any variable in the information set Ψ_{t-1}.

The model parameters are estimated by making use of quasi-maximum likelihood (QMLE). The model building process for STAR models is divided in three separate stages as described in Granger et al. [14] and Teräsvirta [31]. Finally the STAR models are evaluated based on within-sample performance following Eitrheim and Teräsvirta [10].

3.3 MLP model

ANN models are considered to be universal approximators and good forecasters of a wide variety of nonlinear patterns. The models are able to include regime switches as well as other nonlinearities. Based on the findings on the MAE and the Diebold and Mariano test [8], provided in Rodriguez-Perez [29], the choice for the ANN model consist of the MLP model. The MLP model can be seen as a feedforward network, which means that the connections between the

nodes do not form a cycle. A general form of the nonlinear MLP(p, q) model for a single hidden layer network is

$$r_t = \beta_0 + \sum_{j=1}^{q} \beta_j g\Big(\sum_{i=1}^{p} \phi_{ij} r_{t-i} + \phi_{0j} \Big) + \varepsilon_t. \tag{5}$$

Here r_t defines the return at time t or also referred to as the system output. The number of inputs, in this case lagged stock returns, as explanatory variables is given by p. The set of lagged returns then equals $R = (1, r_{t-1}, r_{t-2}, ..., r_{t-p})'$ including a constant term. The parameter vector is $\theta = (\beta', \phi')'$, with $\beta = (\beta_1, ..., \beta_q)'$ and $\phi = (\phi_{1j}, ..., \phi_{pj})'$ for $j = 1, ..., q$ which brings together all the weights of the network, with β_j representing the weights from the hidden unit to the output unit and ϕ_{ij} the weights from the input layer to the hidden unit j. The function $g(\cdot)$ determines the connections between nodes of the hidden layer and is called the hidden unit activation function, which enhances the nonlinearity of the model. More clearly, $g(\cdot)$ can take multiple forms, such as a threshold function, which produces binary (0/1) or ($\pm$1) output, or it can be defined as the sigmoid function, which produces an output between 0 and 1. In (5) ε_t defines the residual which is assumed to be i.i.d. For a more concise overview and interpretation of (5) review Kuan and White [21].

3.4 Hybrid model

The proposed ARIMA, STAR and MLP models have achieved successes in modelling and forecasting in their own linear or nonlinear domains. However none of the individual models can be considered as a suitable universal model for all circumstances. Using a linear ARIMA for modelling complex nonlinear problems may very well be inappropriate. Additionally, STAR and MLP models can have inferior results when used on a clearly linear problem. Denton [7] showed that when outliers or multicollinearity exist within the data, the neural networks do outperform the linear models significantly. Markham and Rakes [26] also found that the performance of MLP models heavily relies on the size of the sample and the noise level. This concludes that it is not always a wise choice to simply apply MLP models to any type of data. Since it is difficult to identify the complete characteristics of a time series dataset, combining both the linear and nonlinear modelling capabilities can be a good strategy.

It can be opportunistic to consider a time series as composed of a linear autocorrelation structure and a nonlinear component. Separating the returns in two parts as follows

$$r_t = L_t + N_t, \tag{6}$$

where L_t denotes the linear part of the returns and N_t represents the nonlinear component. These two components can be estimated by making use of the two models. First, the linear component can be taken account of by modelling an ARIMA model, then the remaining residuals from this linear model will only contain the nonlinear relationship. If we define ε_t as the residual at time t from the linear model, then

$$\varepsilon_t = r_t - \hat{L}_t, \tag{7}$$

where $\hat{L}_t$ resembles the forecasted value for time t of the estimated relationship in (1). The residuals are of importance in the diagnosis to decide whether the linear models are sufficient. A linear model is not sufficient if the residuals still contain linear correlation structures. However, analysis of the residuals is not able to detect any nonlinear patterns in the data. In general there is currently no universal diagnostic statistic for checking nonlinear autocorrelation relationships.

Even if the model passed the linear diagnostic check, the model may still not be adequate in modelling the nonlinear relationships. Any significant nonlinear pattern that exhibit in the residuals will typically indicate the shortcoming of the ARIMA model. By modelling these residuals using a MLP model, nonlinear relationships can be discovered. A MLP model for the residuals with n input nodes is given by

$$\varepsilon_t = f(\varepsilon_{t-1}, \varepsilon_{t-2}, ..., \varepsilon_{t-n}) + \eta_t. \tag{8}$$

The function $f(\cdot)$ is a nonlinear function determined by the neural network and η_t equals the i.i.d. random error at time t. Note that if the function $f(\cdot)$ is inappropriate, the error term does not necessarily have to be random. This underlines the critical importance of correct model identification. The forecast of (8) can be expressed as $\hat{N}_t$. By combining the results the forecast of the returns is of the form

$$\hat{r}_t = \hat{L}_t + \hat{N}_t. \tag{9}$$

Although this hybrid structure could lead to improvements in performance, there is some part of subjective judgement of the model order as well as the model adequacy. For example, suboptimal models may be used in the hybrid model. The current practice of Box-Jenkins method focusses on a low order of autocorrelation. Models are considered adequate if low order autocorrelations are not significantly different from zero, though significant autocorrelations of higher order may still exist. This form of suboptimality may not affect the usefulness of the hybrid model as Granger [13] pointed out that in order to produce superior forecasts, the hybrid model should consist of component models which should be suboptimal.

Concluding, the proposed methodology of the hybrid model consist of mainly two steps. The first step equals fitting an ARIMA model to analyse the linear part of the data. The second step then is to develop an ANN model for the residuals from the fitted ARIMA model. Since the ARIMA model only takes account of the linear structure of the data, the residuals of this model will contain information about the nonlinearity. The results of the ANN model can then be used as predictions of the remaining error terms of the ARIMA model. This hybrid structure combines the linear capabilities of the ARIMA model as well as the ANN model in determining different patterns. It is likely that this hybrid model is advantageous by modelling linear as well as nonlinear patterns separately and then combining the forecasts to improve overall performance.

3.5 Forecasting methods

The produced forecasts are based on one-step-ahead and multi-step-ahead methods. Two notes on forecasting procedures: firstly, though one-step-ahead forecasts are calculated by using the estimated parameters and the actual values for lagged returns, also recursive estimation of the parameters are constructed. This is done by using an expanding window of the sample data and by updating the parameters each period as a new rolling forecast. Secondly, nonlinear multi-step-ahead forecasting is not straightforward. Multi-step forecasts are more difficult to assess than for linear models, because exact analytical solutions are not available most of the times when the innovation distribution is unknown. For linear models this does not matter as the multi-step predictors depend on the parameters and historical data and not on the innovation distribution. Monte Carlo simulation can overcome this problem of unknown innovation distribution for evaluating forecast performance for nonlinear STAR models. For evaluation purposes simulation with resampling from the residuals of the established STAR model with

1000 replications is considered.

For the ANN model forecasts case, mainly one-step-ahead forecasts are considered in the literature. This is due to the fact that there is an increased layer of difficulty in the application of multi-step predictions. The most common method for ANN multi-step forecasting consists of training a predictor for the one-step-ahead forecast and using it in a recurrent way for the corresponding multi-step forecast. By making use of this method, the forecasts provided by the ANN model for the next step are fed back into the input layer of the model until the desired forecast horizon is reached. This method is usually called the iteration prediction. Applying this method requires the same number of data points for training as in the one-step-ahead forecasts.

3.6 Statistical and economic criteria of out-of-sample forecasts

The evaluation of the accuracy of the forecasts and the tests are based on $h = 1, ..., H$ forecasting periods for stock index returns r_h, called $\hat{r}_h$. For performance comparing the linear ARIMA model, nonlinear STAR and ANN models and the hybrid model are considered.

3.6.1 Statistical measures and tests

In order to be able to compare the predictive ability of the different models, multiple out-of-sample forecasts criteria and statistical tests are evaluated. The accuracy of the forecast of the estimated models is compared by making use of the MAE, MAPE, RMSE, Theil's U-statistic and the proportion of the times that the signs of the stock index returns are correctly predicted (Signs). The various hypotheses tests include the Pesaran and Timmermann test [30] for examining directional prediction accuracy of changes. Under the null hypothesis, observed and forecasted values are independent.

Also the forecast encompassing test for out-of-sample forecasts given by Clements and Hendry [6] is incorporated. Under the null hypothesis model i encompasses model j based on the forecast errors from the two models (e_i for model i and e_j for model j). This test uses two regressions, the first involves a least squares regression on $e_{ih} = \alpha_1 + \lambda_1(e_{ih} - e_{jh}) + u_h$, for $h = 1, ..., H$ and obtains estimated coefficient $\hat{\lambda}_1$. The second regression obtains estimated coefficient $\hat{\lambda}_2$ from $e_{jh} = \alpha_2 + \lambda_2(e_{ih} - e_{jh}) + u_h$ for $h = 1, ..., H$. In both cases I analyse whether $\hat{\lambda}_1$ and $\hat{\lambda}_2$ are statistically significant and thus indicate that one model encompasses the other.

Furthermore, the differences between RMSE is checked on significance for out-of-sample forecasts. This is done by making use of the equality of competitive forecasts by the Diebold and Mariano forecast test [8], which is a rather general test of the null hypothesis of equal forecast accuracy between two competing models. Given the one- or multi-step-ahead predictions, and given the corresponding prediction errors of e_{1h} and e_{2h} the null hypothesis of $E[d_h] = 0$ is tested, where $d_h = g(e_{1h}, e_{2h})$ (a function of prediction errors). I choose the function $g(\cdot)$ to be of the absolute type $d_h = |e_{1h}| - |e_{2h}|$.

4 Data

For comparing purposes between the different models I analyse the same dataset as described in Pérez-Rodríguez et al. [29], which includes one of the official indexes of the Madrid Stock Market, the Ibex-35 (P_t). This index is composed of the 35 most liquid assets which are listed in the Computer Assisted Trading System (CATS), which during the analysed period were among the

highest in trading volume in cash pesetas. This index is considered to be highly representative for the market and is fitted by capitalisation and dividends of the assets included, but not by expansions in capital. The index is designed to be used as a reference value in derivatives trading products, e.g., options and futures. The CATS system is introduced onto the Madrid Stock Market in September 1989, whereas my sample period starts in September 1991 due to a lack of data. This study makes use of daily closing prices of the Spanish Ibex-35 index ranging from 10 September 1991 to 10 February 2000, with a total of 2093 observations. This closing prices series is transformed into a logarithmic value to compute continuous returns according to the following expression: $r_t = log(\frac{P_t}{P_{t-1}})$.

5 Model identification and out-of-sample forecast results

5.1 Model specification and estimation

In this section I am going to analyse the out-of-sample forecasting ability of all the individual models and the combined hybrid model. This forecasting ability is based on statistical accuracy and economic criteria and is considered for both one- and multi-step-ahead forecasts. The forecasts produced are compared for three different forecast horizons $h = 1, ..., H$, where $H \in \{100, 200, 500\}$. For the forecast horizon of 500 trading days the European stock exchange crisis of 1998 is included in the forecast period, which makes it particularly suitable for obtaining robust results.

Prior to forecasting, there is an identification and estimation step for the models of interest for in-sample performance. For the ARIMA(p, d, q) model this boils down to coefficient estimation for the in-sample-period of data using different values for the lagged and integrated variables. I choose $p, d = 0$ and $q = 1$ for all the different forecast horizons on the basis of the BIC. The ARIMA(0,0,1) model is equal to a MA(1) model, so a relative simple moving average model with one lag is best able to capture the data for the considered combinations of p, d and q.

The two STAR models are identified and estimated in a similar fashion. This modelling procedure is described in Granger and Teräsvirta [14]. The STAR models also include a lagged component and an added delay parameter $S_t = r_{t-d}$ which is considered over the range $1 \leq d \leq 12$. For all ESTAR models with different time horizons the amount of included lagged variables $p = 1$, again based on the BIC. For LSTAR models I find that the best model is specified for $p = 1$ for the in-sample-period for establishing 100-trading day forecasts, $p = 3$ for the 200-trading day hold-out sample and $p = 2$ for 500-trading day hold-out sample.

By making use of the linearity test I choose $d = 6$ for the 100-trading day hold-out sample, $d = 4$ for the 200-trading day hold-out sample and $d = 1$ for the 500-day hold-out sample. The parameters in the STAR models are estimated by the QMLE method and the BFGS numerical algorithm.

Lastly, for the MLP model, I include one hidden layer. In general one hidden layer is specified, because this is generally effective for capturing nonlinear structures [1]. For the hidden layer various hidden units or elements are considered, as denoted by q. The parameters p, q and the activation function $g(\cdot)$ are not chosen upfront. The amount of lags p are calculated by means of a sequential validation, so models with different combinations of values for p and q are estimated. The relevant rank of the parameters that is taken into account is $p, q = \{1, ..., 5\}$. The model consists of lagged stock returns which are scaled assuming a uniform distribution within

the interval [-1,1], this is done for speed improvements and to make sure that the algorithm does not get stuck in local optima. For the hidden unit activation function $g(\cdot)$ the hyperbolic tangent function is used, which provides a better fit over other activation functions [16].

Afterwards, the MLP model is trained over a given period using 1500 cycles and cross-validation. This set is used to estimate the weights of the neural network. For improvements of the in-sample fitting, the estimated set of weights are used as a set of initial values. Furthermore cross-validation is performed to avoid overfitting of the data. The models for different combinations of p and q are then analysed in terms of minimisation of the RMSE. Best models are yielded for $p = 2$ and $q = 4$ for all training samples with different forecast horizons.

Once the models are identified and the parameters are estimated, the forecasts can be constructed. The findings of this empirical research are discussed in the next sections.

5.2 Statistical analysis of out-of-sample forecasts

Here I will discuss the model selection procedures which are described in section 3.6.1 focussing on the the individual models and the combined hybrid model. First of all the basic statistic properties of the produced forecasts are evaluated together with the market timing ability of the different models. This is done for all different forecast methods and horizons. Also the hypothesis of Pesaran and Timmermann [30] of directional accuracy is tested. Furthermore, the forecast encompassing test is evaluated for all forecasts and finally the Diebold-Mariano test for equal accuracy is compared for the predictions.

5.2.1 Goodness of forecast results

Table 1 shows the results of the different statistical measures for the forecasts over the different procedures and horizons. Next to that the Signs-value and P-values of the Pesaran-Timmermann test for directional accuracy (DA) are given for all models. The main findings of the forecast evaluation are that even with inclusion of the hybrid model no single model performs best if evaluated over all different forecasts. In fact, it is noticed that the results are fairly heterogeneous over the different models. The simple ARIMA model provides on average pretty decent forecasts, but those remain somewhat conservative and do not capture the peaks in the returns. In general the LSTAR and ESTAR models perform reasonably well in terms of the MAE and MAPE, especially for the longer horizon (recursive) one-step-ahead forecasts. As in Pérez-Rodríguez et al. [29] it is found that the MLP model provides good forecast for the short-horizon forecasts, but deteriorates when the forecast horizon is extended. In terms of MAPE the MLP and hybrid model do not seem to capture the returns very well, since these values increase rapidly as horizon increases over all forecast methods.

Nevertheless, the MLP model and the hybrid model are of added value. Especially for the short time horizon forecast, the hybrid model seems particularly useful. The hybrid model surpasses the nondrift random walk method as calculated in Theil's U-statistic at all forecast horizon and forecast method. This means that the Ibex-35 stock index return can be predicted by these models and that the random walk model does not seem to provide a reasonable description of the movement of the index return. Next to that, if we look at it from a market timing perspective the hybrid model gives a slight improvement with respect to the other (non)linear models. This is seen by the relatively high Signs values. However, this improvement is not significant most of the times, since we do not reject the null hypothesis of the DA test that forecasts and realisations are independent. Only for the 200-day horizon one-step-ahead forecast this is

rejected at a 5% significance level. This shows that the hybrid neural network model does not perform significantly better than the other (non)linear models, because the null hypothesis of independent realisations and forecasts is not often rejected.

Concluding, the hybrid model can be useful in predicting the sign of the returns, with relative low errors for short forecast horizons. Also in terms of surpassing the nondrift random walk the hybrid model seems superior. However, it is remarkable that the performance over longer horizons is inferior to the individual (non)linear models in terms of forecast error.

Table 1: Forecast evaluation statistics and DA test for all models

	H=100						H=200						H=500					
Models	MAE	MAPE	RMSE	Theil	Signs	DA	MAE	MAPE	RMSE	Theil	Signs	DA	MAE	MAPE	RMSE	Theil	Signs	DA
One-step-ahead																		
ARIMA	0.9796*	102.53	1.2005*	0.8586	0.52	0.69	0.8987	127.10	1.1323	0.8743	0.515	0.93	1.2348*	149.61	1.6761*	0.8757	0.532	0.57
ESTAR	0.9833	103.46	1.2037	0.8574	0.53	0.89	0.8965	116.04*	1.1130*	0.8634	0.535	0.84	1.2405	133.59*	1.6955	0.8241	0.542*	0.18
LSTAR	0.9834	101.26*	1.2051	0.8780	0.53	0.87	0.8986	118.01	1.1299	0.8746	0.49	0.78	1.2585	140.74	1.7153	0.7683*	0.54	0.20
MLP	0.9828	112.09	1.2157	0.8056	0.56*	0.45	0.9107	159.81	1.1525	0.8072	0.53	0.57	1.2954	228.91	1.7443	0.7943	0.50	0.66
Hybrid	0.9853	110.37	1.2065	0.7915*	0.56*	0.47	0.8915*	162.97	1.1210	0.7809*	0.58*	0.04^a	1.2867	249.27	1.7434	0.7807	0.52	0.46
Recursive one-step-ahead																		
ARIMA	0.9805	102.38	1.2010*	0.8617	0.52	0.69	0.8986	125.85	1.1324	0.8783	0.515	0.93	1.2344	146.61	1.6761*	0.8809	0.53	0.59
ESTAR	1.0014	99.84*	1.2238	0.9304	0.58	0.90	0.9004	116.13	1.1320	0.9205	0.525	0.89	1.2314*	119.28	1.6884	0.9572	0.526	0.86
LSTAR	1.0015	99.97	1.2245	0.9357	0.56	0.85	0.8931*	109.86*	1.1255*	0.9069	0.565*	0.78	1.2315	112.31*	1.6888	0.9449	0.526	0.30
MLP	0.9838	109.22	1.2145	0.8111	0.56	0.51	0.9171	153.27	1.1506	0.8212	0.52	0.57	1.2602	192.56	1.7095	0.8229	0.514	0.47
Hybrid	0.9761*	110.82	1.2149	0.8056*	0.6*	0.38	0.9083	157.26	1.1381	0.8035*	0.555	0.85	1.2578	199.07	1.6859	0.8212*	0.534*	0.62
Multi-step-ahead																		
ARIMA	0.9966	98.97	1.2199	0.9433	0.61*	-	0.8985	109.90	1.1318	0.9392	0.545	-	1.2305	113.36	1.6870*	0.9609	0.546	-
ESTAR	0.9964	98.01*	1.2226	0.9364	0.61*	-	0.8965*	113.49	1.1289*	0.9256	0.565*	-	1.2334	111.25*	1.6878	0.9624	0.54	-
LSTAR	0.9977	99.24	1.2207	0.9318	0.61*	-	0.8986	108.29*	1.1327	0.9250	0.555	-	1.2326	113.68	1.6877	0.9627	0.534	-
MLP	0.9631*	114.52	1.2259	0.7581	0.61*	-	0.9031	155.00	1.1417	0.8312	0.545	-	1.2283*	145.84	1.6894	0.9071	0.548*	-
Hybrid	0.9750	102.23	1.2117*	0.7447*	0.61*	-	0.9071	166.56	1.1469	0.8133*	0.545	-	1.2592	254.85	1.7325	0.8005*	0.548*	-

The table reports the goodness of forecast and the DA test of alternative forecasting models. P-values appear for the DA test. * indicates that the result is the best between models in the same column. ^{a} indicates a significant P-value of the DA test at a 5% significance level. No P-values for the DA test in the multi-step case are given because all models take on positive values and tend to be constant.
Note: All MAE and RMSE values should be multiplied by 10^{-2}

5.2.2 Forecast encompassing results

Table 2 reports the findings of the forecast encompassing test, which is a test for a two-by-two comparison of the competing models. The table shows the P-values of t-statistics on the null hypothesis of $\hat{\lambda}_1 = 0$ and $\hat{\lambda}_2 = 0$ for all $i - j$ comparisons (for $i, j = \{ARIMA, ESTAR, LSTAR, MLP, hybrid\}$). The general element is denoted by b_{ij}, where i is the i-th row and j is the j-th column. When referred to $b_{ESTAR,ARIMA}$ the P-value for $\hat{\lambda}_1$ (lower triangular matrix) is considered and when referred to $b_{ARIMA,ESTAR}$ (upper triangular matrix) the P-value for $\hat{\lambda}_2$ is considered.

From table 2 two main results stand out. Firstly, the forecast encompassing test shows that the ARIMA model almost always encompasses the other models for the longest horizon for the (recursive) one-step-ahead forecasts. The forecasts of the ARIMA model are conditionally more efficient than those of the other models for the forecast horizon $H = 500$. Given that $b_{i,ARIMA} < 0.05$ for $i = \{ESTAR, LSTAR, MLP, hybrid\}$ for the one-step-ahead and also holds for $i = \{ESTAR, hybrid\}$ for the recursive one-step-ahead forecasts. This indicates that $\hat{\lambda}_1$ is statistically significant (P-value lower than 0.05 in lower triangular matrix). Thus, we reject the null hypothesis and the ARIMA model encompasses the other models at a 5% significance level. Secondly, one can see that the hybrid model does not often surpass the other individual models. In fact, it is shown in the table that the hybrid model is more often being encompassed than encompassing other models. The times that the hybrid model is encompassed is given when $b_{hybrid,j} < 0.05$ for $j = \{ARIMA, ESTAR, LSTAR, MLP\}$ and the time it encompasses another model is given when $b_{i,hybrid} < 0.05$ for $i = \{ARIMA, ESTAR, LSTAR, MLP\}$. Especially at the 500-trading day forecast for (recursive) one-step-ahead methods, the other models are conditionally more efficient than the proposed hybrid model. On the other hand, the opposite holds for the 200-trading day one-step-ahead forecasts, where the hybrid model significantly encompasses every other model at a 5% significance level.

The reason behind the poor 500-day forecast performance of both the MLP as well as the hybrid model is most likely due to the fact that overfitting occurs. The 500-trading day forecasts includes the European stock exchange crisis which generated high volatility across several European markets. This volatility is not captured well by the specified (neural network) models. A solution to overcome this overfitting is to make use of an optimized approximation algorithm as given by Liu et al. [23]. This algorithm is proposed to solve overfitting by utilizing a quantitative stopping criterion based on the signal-to-noise-ratio figure (SNRF). The SNRF estimates ratios of expected prediction errors and based on this figure an optimized approximation algorithm can be proposed. Overfitting is automatically detected by the algorithm by only making use of the training errors. The algorithm itself has been validated for optimizing the number of hidden neurons for MLP models and can often increase forecasting power [23].

Nonetheless, the table also shows that in some cases the hybrid model does encompass other models. Mainly this occurs in the case of the 200-trading day one-step-ahead forecast. This indicates that for some forecasting methods and horizons, the hybrid model can be slightly better in terms of conditional efficiency. Especially during periods of relatively constant volatility the hybrid model encompasses the other models as both linear as nonlinear parts are captured adequately.

Therefore I conclude that the hybrid model can serve as an attractive alternative model for forecasting in some cases. However, it does not consistently produce superior forecasts in

general. Another remarkable finding is that for the multi-step-ahead case only one model is encompassed, whereas Pérez-Rodríguez et al. [29] found much more rejections of the null hypothesis for long horizon multi-step-ahead horizons. This difference in P-values can possibly be explained by the selection of the training period, as the sample period differs with a lack of two years, which yields different model parameters and thus different forecasts.

Table 2: P-values associated with the heteroskedasticity-robust t-statistics on $\hat{\lambda}_1$ and $\hat{\lambda}_2$ for all possible i-j comparisons based on OLS regressions

Models	H=100						H=200						H=500				
	ARIMA	ESTAR	LSTAR	MLP	Hybrid		ARIMA	ESTAR	LSTAR	MLP	Hybrid		ARIMA	ESTAR	LSTAR	MLP	Hybrid
One-step-ahead																	
ARIMA	-	0.23	0.26	0.35	0.52	ARIMA	-	0.40	0.35	0.37	0.02*	ARIMA	-	0.69	0.32	0.87	0.83
ESTAR	0.17	-	0.74	0.24	0.44	ESTAR	0.88	-	0.55	0.49	0.03*	ESTAR	0.00*	-	0.41	0.08	0.10
LSTAR	0.15	0.49	-	0.13	0.40	LSTAR	0.80	0.62	-	0.58	0.04*	LSTAR	0.00*	0.00*	-	0.00*	0.00*
MLP	0.79	0.67	0.41	-	0.62	MLP	0.89	0.53	0.56	-	0.03*	MLP	0.00*	0.00*	0.00*	-	0.65
Hybrid	0.16	0.19	0.23	0.11	-	Hybrid	0.29	0.22	0.26	0.23	-	Hybrid	0.00*	0.00*	0.00*	0.00*	-
One-step-ahead (recursive regression)																	
ARIMA	-	0.40	0.37	0.62	0.58	ARIMA	-	0.31	0.11	0.06	0.26	ARIMA	-	0.93	0.95	0.16	0.10
ESTAR	0.04*	-	0.86	0.11	0.12	ESTAR	0.36	-	0.13	0.35	0.15	ESTAR	0.01*	-	0.68	0.00*	0.00*
LSTAR	0.04*	0.74	-	0.10	0.12	LSTAR	0.70	0.93	-	0.30	0.29	LSTAR	0.06	0.50	-	0.00*	0.00*
MLP	0.27	0.81	0.75	-	0.39	MLP	0.00*	0.00*	0.00*	-	0.62	MLP	0.82	0.89	1.00	-	0.11
Hybrid	0.17	0.47	0.51	0.24	-	Hybrid	0.07	0.04*	0.02*	0.85	-	Hybrid	0.00*	0.01*	0.01*	0.00*	-
multi-step-ahead																	
ARIMA	-	0.61	0.70	0.54	0.60	ARIMA	-	0.21	0.90	0.86	0.90	ARIMA	-	0.77	0.88	0.00*	0.72
ESTAR	0.41	-	0.71	0.40	0.48	ESTAR	0.45	-	0.54	0.42	0.46	ESTAR	0.47	-	0.68	0.42	0.46
LSTAR	0.50	0.94	-	0.49	0.58	LSTAR	0.53	0.19	-	0.57	0.52	LSTAR	0.57	0.84	-	0.51	0.55
MLP	0.98	0.60	0.69	-	0.59	MLP	0.78	0.20	0.96	-	0.46	MLP	0.76	0.73	0.83	-	0.75
Hybrid	0.32	0.71	0.79	0.54	-	Hybrid	0.98	0.22	0.90	0.48	-	Hybrid	0.74	0.76	0.86	0.68	-

Two triangular matrices for each horizon and forecast method. Upper triangular matrix corresponds to P-values for $\hat{\lambda}_2$ and lower triangular matrix corresponds to P-values for $\hat{\lambda}_1$. When I refer to the P-value for $\hat{\lambda}_1$, the row is the i-th model and the column is the j-th model. When I refer to the P-value for $\hat{\lambda}_2$, the row is the j-th model and the column is the i-th model. * indicates that $\hat{\lambda}_1$ or $\hat{\lambda}_2$ is significant at a 5% significance level.

5.2.3 Diebold and Mariano test results

Also I compare the produced forecasts by means of forecast accuracy by making use of the Diebold and Mariano (DM) test considering the MAE loss function, which table 3 shows. The test examines whether there is a significant difference in the MAE of the forecasts obtained from the different models. The performed DM test is based on the absolute difference between the MAE of two models, i.e. $d_h = |e_{1h}| - |e_{2h}|$. The P-value associated with this outcome is given in the upper triangular matrix for each combination of models over each horizon. The lower triangular matrix is omitted because the matrix is diagonal symmetric, but with a multiplication of -1 of the upper triangular matrix. The e_2 forecast errors are shown in the columns and the e_1 errors in the rows.

In the table multiple findings stand out. For the long horizon one-step-ahead forecasts, the MAE of both the MLP and the hybrid model are significantly larger than those of the ARIMA model. This indicates that the ARIMA model is relatively good at predicting the stock index returns compared to (combined) neural network models over longer horizons. This holds for the 500-trading day one-step-ahead forecast as well as the 200-trading day recursive forecasts for the MLP model. The MLP model produces significantly inferior forecasts and does not seem to capture the movement of the stock index out-of-sample.

For the short forecast horizon of 100 trading days, the MLP and hybrid model perform well over all three forecast methods. However, the MAE is not significantly lower than other models at the conventional 5% significance level, but is in some cases at the 10% significance level. Concluding, the hybrid model seem to capture the stock index return fairly well for short forecast horizons of 100- and 200-trading days. However, as time horizon grows, the forecasts reduce in terms of forecast accuracy. Also table 3 shows evidence of no superior model or forecast method in general for the sample.

Table 3: Diebold-Mariano test statistic for $d_h = |e_{1h}| - |e_{2h}|$ (MAE, upper triangular matrix in each block)

Models	h=100						h=200						h=500				
	ARIMA	ESTAR	LSTAR	MLP	Hybrid		ARIMA	ESTAR	LSTAR	MLP	Hybrid		ARIMA	ESTAR	LSTAR	MLP	Hybrid
One-step-ahead																	
ARIMA	-	-1.8	-1.0	0.9	-0.2	ARIMA	-	0.6	0.0	-0.2	0.4	ARIMA	-	-0.5	-1.5	-2.3*	-2.6*
ESTAR	-	-	0.0	1.2	-0.1	ESTAR	-	-	-0.3	-0.4	0.3	ESTAR	-	-	-1.5	-1.7	-2.1*
LSTAR	-	-	-	1.3	-0.1	LSTAR	-	-	-	-0.2	0.4	LSTAR	-	-	-	-0.8	-1.1
MLP	-	-	-	-	-0.5	MLP	-	-	-	-	0.5	MLP	-	-	-	-	-0.3
Hybrid	-	-	-	-	-	Hybrid	-	-	-	-	-	Hybrid	-	-	-	-	-
One-step-ahead (recursive regressions)																	
ARIMA	-	-1.5	-1.5	0.6	0.2	ARIMA	-	-0.2	0.5	-3.1*	-0.6	ARIMA	-	0.3	0.3	0.0	-1.6
ESTAR	-	-	0.0	1.6	1.0	ESTAR	-	-	0.9	-2.5*	-0.4	ESTAR	-	-	0.0	-0.3	-1.5
LSTAR	-	-	-	1.6	1.0	LSTAR	-	-	-	-2.7*	-0.8	LSTAR	-	-	-	-0.3	-1.6
MLP	-	-	-	-	0.0	MLP	-	-	-	-	1.7	MLP	-	-	-	-	-1.5
Hybrid	-	-	-	-	-	Hybrid	-	-	-	-	-	Hybrid	-	-	-	-	-
Multi-step-ahead																	
ARIMA	-	0.1	-0.3	1.9	1.5	ARIMA	-	0.7	0.0	-1.1	-0.6	ARIMA	-	-1.8	-0.8	-1.0	-1.6
ESTAR	-	-	-0.2	1.2	1.5	ESTAR	-	-	-0.5	-1.2	-0.8	ESTAR	-	-	0.7	-0.7	-1.4
LSTAR	-	-	-	1.8	1.7	LSTAR	-	-	-	-0.7	-0.6	LSTAR	-	-	-	-0.8	-1.5
MLP	-	-	-	-	1.4	MLP	-	-	-	-	-0.3	MLP	-	-	-	-	-3.7*
Hybrid	-	-	-	-	-	Hybrid	-	-	-	-	-	Hybrid	-	-	-	-	-

Critical distribution values for $N(0,1)$ is 1.96 at 5% significance level. * indicates that the MAE of one of the models is significantly different from the other model. Values smaller than zero indicate that forecasts produced by model i are lower than those of model j. The opposite holds for values which are greater than zero.

6 Conclusions

I analysed five different models for forecasting the out-of-sample daily returns of the Ibex-35 index. The models of interest are the linear ARIMA model, the nonlinear ESTAR and LSTAR smooth transition autoregressive models, an artificial neural network MLP model and a combined hybrid ARIMA-MLP model. The comparison between these forecasts is carried out on basis of a set of statistical criteria, while making use of different forecast methods and multiple forecast horizons in order to obtain robust results.

As shown in Pérez-Rodríguez et al. [29], the random walk model is not a reasonable description of the movement of the Ibex-35 stock index returns. This indicates a rejection of the efficient market hypothesis and argues that nonlinear models may be more appropriate. Pérez-Rodríguez et al. suggest that ANN techniques may be a more appropriate direction for improvement of forecasting ability. This seems to hold for some cases of the one-step-ahead forecasts. For further improvement I propose a combined ARIMA-MLP model. This model is able to capture both the linear and the nonlinear habits of the stock index returns. This property could be of potential added benefit over the individual MLP and ARIMA models.

Empirical analysis of the statistical properties of the produced forecasts shows that the hybrid model can be particularly useful for short forecast horizons of 100- and 200-trading days, but deteriorates as horizon grows. Also, the hybrid model is not always able to deliver superior forecasts over the other proposed methods. I find that the ARIMA and STAR models produce superior forecasts for 500-trading days. Over long horizons the forecast error increases and MAPE explodes. Despite the fact that the market timing ability of the hybrid model is not always significant, the hybrid model does a better job at correctly predicting the sign of the index return than the other models do. This also holds for longer forecast horizons.

This pattern of fruitful short horizon forecasts, but inferior long horizon forecasts also resides in the forecast encompassing test. For one-step-ahead forecasts and at short horizons the hybrid model encompasses several other individual models, but at longer horizons all other individual models encompass the hybrid model.

Furthermore this holds for the Diebold and Mariano test for different absolute forecast errors as well. For small horizons it can not be rejected that the hybrid model produces equal forecasts in terms of MAE. However for longer horizons, it is rejected with a 5% significance level that the hybrid model produces equal forecasts, due to relatively high MAE values at expense of the hybrid model. The inclusion of the crash of European stock indices shows that the hybrid model is not able to produce adequate forecasts for this period of unforeseen high volatility.

After assessing and comparing different statistical criteria, I conclude that the return on the Ibex-35 index can be predicted well by making use of a hybrid ARIMA-MLP model for a relatively short time horizon, which is also supported by Zhang [37]. However, these findings do not persist for longer time horizons with high volatility. Therefore, the hybrid model can deliver a slight improvement in investment strategies for short-period forecasting over individual (non)linear models.

7 Discussions

In my empirical research the performance of the proposed hybrid ARIMA-MLP model is solely evaluated for one stock index return series, this lead to limitations of the possible forecast improvement of the hybrid model. A possibly interesting research direction is to investigate the performance of the hybrid model for aggregated stock returns or bonds as these have different, possibly correlated, movements which the model might be capable of to account for.

Besides that, the proposed model is also heavily dependent on the period that is used for model training. Multiple training periods, or longer periods with cross-validation could make sure that no overfitting occurs in-sample and better forecasts can be made in different states of the economy. To improve longer horizon forecasting overfitting needs to reduce. Due to the fact that a European stock exchange crash is included in the longer horizon, the benefit of a (hybrid) neural network model is marginal. The proposed model is not able to handle periods of volatility clustering well within the stock index return data. I propose a method in section 5.2.2, as used by Liu et al. [23], which can overcome overfitting and may be valuable for producing less training-period dependent forecasts.

In this paper I make use of the hybrid ARIMA-MLP model as proposed by Zhang [37]. However, Khashei and Bijari [19] suggest a slight modification which seem to benefit forecasting purposes. Their model also consists of a sum of a linear and a nonlinear component, but handles the produced errors of the ARIMA model somewhat different. Instead of approaching a suitable fit by making use of only the lagged errors, also past original return values, present ARIMA-forecasted return values and past error sequence values are given as input to the ANN model. This method shows to have better performance than Zhangs method in various applications in terms of accuracy of the forecasts. Whether this method could benefit stock index returns forecasting remains for further research.

Next to that, not only observed or predicted values for the stock index, but also macroeconomic variables could be of great influence on predicting the movement of the stock index returns. Changes in GDP, incomes or the housing market may heavily effect the underlying stocks in the index and have additional explanatory power when forecasting the index. Adding these variables as input for (hybrid) neural network models may yield interesting results.

Furthermore the choice of the activation function and the the rescaling of the input data remains for discussion. In this research I consider the relationship between one input and one neuron to be defined by a hyperbolic tangent activation function. However, Kalman and Kwasny [18] discuss whether the hyperbolic tangent function should be used instead of for example the sigmoid function. On the other hand, empirical results obtained by Maier and Dandy [24] indicate that not only the hyperbolic tangent function is preferred over the logistic sigmoid function in terms of computation speed, but also in forecast performance. However, my study lacks a comparison between the performance of the logistic sigmoid function and the hyperbolic tangent function.

Lastly, the rescaling of the inputs to [-1, 1] limits the output range of the activation function approximately to [-0.7616, 0.7616]. These bounds are far away from the extreme limits of the activation function. A small range of output values will make the output less sensitive to the change of the weights between the hidden layer and the output layer, which will make the training process more difficult. On the other hand, because the neurons in MLP models are linearly combinations with many weights, any rescaling of the input vector can be effect-

ively offset by changing the corresponding weights and biases. This gives reasonable doubt whether standardization of the input data may be a better choice than to rescale them into a small interval, especially when the data size is large enough to show presence of extreme returns.

References

[1] Monica Adya and Fred Collopy. "How effective are neural networks at forecasting and prediction? A review and evaluation". In: *J. Forecasting* 17 (1998), pp. 481–495.

[2] Sandy D Balkin and J Keith Ord. "Automatic neural network modeling for univariate time series". In: *International Journal of Forecasting* 16.4 (2000), pp. 509–515.

[3] William G Baxt. "Improving the accuracy of an artificial neural network using multiple differently trained networks". In: *Neural Computation* 4.5 (1992), pp. 772–780.

[4] Anil K Bera and Matthew L Higgins. "ARCH models: properties, estimation and testing". In: *Journal of economic surveys* 7.4 (1993), pp. 305–366.

[5] George EP Box, Gwilym M Jenkins, Gregory C Reinsel and Greta M Ljung. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[6] Michael P Clements and David F Hendry. "On the limitations of comparing mean square forecast errors". In: *Journal of Forecasting* 12.8 (1993), pp. 617–637.

[7] Emily L Denton, Wojciech Zaremba, Joan Bruna, Yann LeCun and Rob Fergus. "Exploiting linear structure within convolutional networks for efficient evaluation". In: *Advances in neural information processing systems*. 2014, pp. 1269–1277.

[8] Francis X Diebold and Robert S Mariano. "Comparing predictive accuracy". In: *Journal of Business & economic statistics* 20.1 (2002), pp. 134–144.

[9] Zhuanxin Ding, Clive WJ Granger and Robert F Engle. "A long memory property of stock market returns and a new model". In: *Journal of empirical finance* 1.1 (1993), pp. 83–106.

[10] Øyvind Eitrheim and Timo Teräsvirta. "Testing the adequacy of smooth transition autoregressive models". In: *Journal of Econometrics* 74.1 (1996), pp. 59–75.

[11] Fernando Fernández-Rodríguez, Simón Sosvilla-Rivero and Maria Dolores Garcia-Artiles. "Dancing with bulls and bears: Nearest-neighbour forecasts for the Nikkei index". In: *Japan and the World Economy* 11.3 (1999), pp. 395–413.

[12] Jan G de Gooijer and Kuldeep Kumar. "Some recent developments in non-linear time series modelling, testing, and forecasting". In: *International Journal of Forecasting* 8.2 (1992), pp. 135–156.

[13] Clive WJ Granger. "Invited review combining forecasts, twenty years later". In: *Journal of Forecasting* 8.3 (1989), pp. 167–173.

[14] Clive WJ Granger, Timo Teräsvirta et al. "Modelling non-linear economic relationships". In: *OUP Catalogue* (1993).

[15] Claas de Groot and Diethelm Würtz. "Analysis of univariate time series with connectionist nets: A case study of two classical examples". In: *Neurocomputing* 3.4 (1991), pp. 177–192.

[16] John P Guiver and Casimir C Klimasauskas. "Applying neural networks, Part IV: improving performance". In: *PC AI Magazine* 5 (1991), pp. 34–41.

[17] Melvin J Hinich and Douglas M Patterson. "Evidence of nonlinearity in daily stock returns". In: *Journal of Business & Economic Statistics* 3.1 (1985), pp. 69–77.

[18] Barry L Kalman and Stan C Kwasny. "Why tanh: choosing a sigmoidal function". In: *International Joint Conference on Neural Networks, 1992. IJCNN.* Vol. 4. IEEE. 1992, pp. 578–581.

[19] Mehdi Khashei and Mehdi Bijari. "A novel hybridization of artificial neural networks and ARIMA models for time series forecasting". In: *Applied Soft Computing* 11.2 (2011), pp. 2664–2675.

[20] Mehdi Khashei, Mehdi Bijari and Gholam Ali Raissi Ardali. "Improvement of auto-regressive integrated moving average models using fuzzy logic and artificial neural networks (ANNs)". In: *Neurocomputing* 72.4-6 (2009), pp. 956–967.

[21] Chung-Ming Kuan and Halbert White. "Artificial neural networks: An econometric perspective". In: *Econometric reviews* 13.1 (1994), pp. 1–91.

[22] Alan Lapedes and Robert Farber. *Nonlinear signal processing using neural networks: Prediction and system modelling.* Tech. rep. 1987.

[23] Yinyin Liu, Janusz A Starzyk and Zhen Zhu. "Optimized approximation algorithm in neural networks without overfitting". In: *IEEE transactions on neural networks* 19.6 (2008), pp. 983–995.

[24] Holger R Maier and Graeme C Dandy. "The effect of internal parameters and geometry on the performance of back-propagation neural networks: an empirical study". In: *Environmental Modelling & Software* 13.2 (1998), pp. 193–209.

[25] Spyros Makridakis. "Why combining works?" In: *International Journal of Forecasting* 5.4 (1989), pp. 601–603.

[26] Ina S Markham and Terry R Rakes. "The effect of sample size and variability of data on the comparative performance of artificial neural networks and regression". In: *Computers & operations research* 25.4 (1998), pp. 251–263.

[27] Paul Newbold and Clive WJ Granger. "Experience with forecasting univariate time series and the combination of forecasts". In: *Journal of the Royal Statistical Society. Series A (General)* (1974), pp. 131–165.

[28] Franz C Palm and Arnold Zellner. "To combine or not to combine? Issues of combining forecasts". In: *Journal of Forecasting* 11.8 (1992), pp. 687–701.

[29] Jorge V Pérez-Rodríguez, Salvador Torra and Julián Andrada-Félix. "STAR and ANN models: forecasting performance on the Spanish "Ibex-35" stock index". In: *Journal of Empirical Finance* 12.3 (2005), pp. 490–509.

[30] Hashem M Pesaran and Allan Timmermann. "A simple nonparametric test of predictive performance". In: *Journal of Business & Economic Statistics* 10.4 (1992), pp. 461–465.

[31] Timo Teräsvirta. "Specification, estimation, and evaluation of smooth transition autoregressive models". In: *Journal of the american Statistical association* 89.425 (1994), pp. 208–218.

[32] Fang-Mei Tseng, Hsiao-Cheng Yu and Gwo-Hsiung Tzeng. "Combining neural network model with seasonal time series ARIMA model". In: *Technological Forecasting and Social Change* 69.1 (2002), pp. 71–87.

[33] Xiping Wang and Ming Meng. "A Hybrid Neural Network and ARIMA Model for Energy Consumption Forcasting." In: *JCP* 7.5 (2012), pp. 1184–1190.

[34] Halbert White. "Economic prediction using neural networks: The case of IBM daily stock returns". In: *IEEE 1988 International Conference on Neural Networks* (1988).

[35] Robert L Winkler. “Combining forecasts: A philosophical basis and some current issues”. In: *International Journal of Forecasting* 5.4 (1989), pp. 605–609.

[36] Peter G Zhang. “A neural network ensemble method with jittered training data for time series forecasting”. In: *Information Sciences* 177.23 (2007), pp. 5329–5346.

[37] Peter G Zhang. “Time series forecasting using a hybrid ARIMA and neural network model”. In: *Neurocomputing* 50 (2003), pp. 159–175.