Quantitative and intelligent methods for economic forecasts (birth rate time series (Ntl))

Katerina ZELA,
Mediterranean University of Albania, Department of Information Technology, Tirana, Albania
katerina.male@umsh.edu.al / male.katerina@yahoo.com

Abstract
The models proposed in the field of time series forecasting are numerous. However, accurate forecasting is still one of the main challenging problems faced by decision makers in various fields, especially in financial markets. Anticipatory and intelligent methods play a crucial role in economic forecasting, helping to increase the accuracy and reliability of forecasts. This is the main reason that many researchers have in the focus of their research the development of methods for improving the performance of forecasts. Using only a linear or non-linear method is not enough to recognize and model all features and functions of the data. In addition to quantitative techniques as well as intelligent techniques that have been developed recently as a result of technological developments, a new technique or hybridization of these two techniques behaves as a very efficient approach. For this reason, in the study it is proposed to improve the modeling and forecasting results by combining a linear A RIMA method with an intelligent non-linear method, the artificial neural network - ANN. Several ARIMA-ANN hybrid model architectures are known, but in this thesis a new A RIMA-ANN technique with improving features of the existing hybrid methods is proposed. The time series under study are the birth rate in Albania obtained by INSTAT and the Bank of Albania with monthly frequencies. Time series analysis goes through three phases: defining characteristics, modeling and forecasting. The analysis is performed with ARIMA, ANN and the proposed ARIMA-ANN method. Data mining, sentiment analysis, and natural language processing (NLP) techniques are used to analyze unstructured data sources. The models, either quantitative or intelligent, give satisfactory results, but the hybrid model shows much higher performance in all series. From the measurement indicators of the prediction performance, the hybrid method has resulted with high efficiency. The ARIMA-ANN model produced much better results than the ARIMA and ANN models. These results are a consequence of the combination of the two techniques and from which results a good recognition of the different structures of the data modes. This study contributes to the field of time series forecasts in Albania and intelligent or quantitative modeling by developing a new forecast model, that of hybridization.

Keywords: time series, modeling, forecasting, ARIMA, ANN, hybrid ARIMA-ANN model.

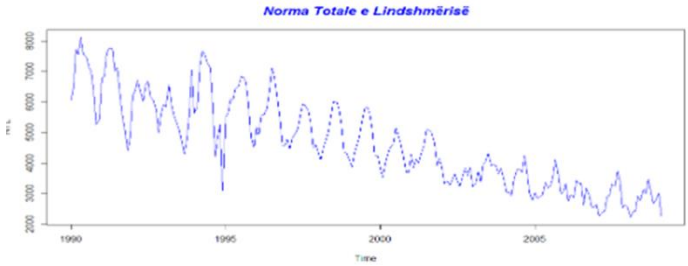
1. Introduction. Descriptive Analysis.
The first series taken into consideration for analysis is the time series of the total birth rate (N TL). The total fertility rate indicates the average number of children in relation to each woman who has the ability to reproduce. This indicator is an average value since some women can give birth to more, some less children and some may not give birth [1]. Taking into consideration the fact that in every marriage a woman is responsible for giving birth and in every marriage there is a high possibility that the woman will give birth and survive the birth of the child, the birth rate has been estimated as the ratio of the number of births per month and the number of marriages per month [2].

NTL= Number of births per month / Number of marriages per month

The NTL is a more direct indicator of the fertility rate than the crude birth rate. This indicator PRESENT the potential for change THE POPULATION OF THE A the country The coefficient of two child ABOUT woman is considered HOW coefficient The

Replacements ABOUT A population, creating A ENDURANCE IN to [3]. Coefficients with above two child show GROWTH THE number THE POPULATION AND Falling THE AGE average. norms UNDER two child show Falling THE number THE POPULATION AND aging THE her _ IN world, in overall, the birth rate is coming up IN Falling AND this trend is MORE The noted IN industrialized countries, where _ spredicted. THAT Number The Population THE line decreasing _ ORDER THE emphasized next years. Compared to the share other of Europe birth rates _ IN Albania HAVE we were THE HIGHER THE AT LEAST UP IN in 2001. Meanwhile that, before 1990s one _ woman birth 6.8 children on average. amendments THE visible occurred after the changes POLITICAL AND ECONOMIC IN in 1990. WHEREAS that, rates AND values traditional they continue THE AFFECTING IN family formation, are _ socio- economic changes those THAT determine MORE too the birth rate. 5.1.2 Graphical analysis of the data IN in relation to the birth rate, the NTL variable (rate total birth rate), are TAKE BY INSTAT database with frequencies monthly IN Albania ABOUT A period FROM from 1990 to in 2013 [4].

More DOWN given Graph ABOUT THE Front PERFORMANCE tall YEARS THE THIS series IN Albania. Graph 5.1Graph i PROGRESS OF THE RATE the totals fertilityGraphically a downward trend is observed seems like stabilize recent years (2006 to continuation).



Graph 1. The graph The PROGRESS OF THE RATE the totals fertility

Graphically a downward trend is observed seems like stabilize recent years (2006 to continuation).

Norms of fertility IN Albania HAVE HAD A level THE up IN the 90s with almost 6 children. ABOUT woman AND COMES IN values MORE THE LOW IN recent years falling _ _ DOWN RATE OF THE substitution which _ there are 2 children for each woman THE FIT ABOUT THE born. As possible tsihetnga chart, fertility series shows a downward trend AND One-component seasonal presented _ BY regular fluctuation _ [5] . Under development a complete analysis of the series, is used functionary in R for its decomposition. _

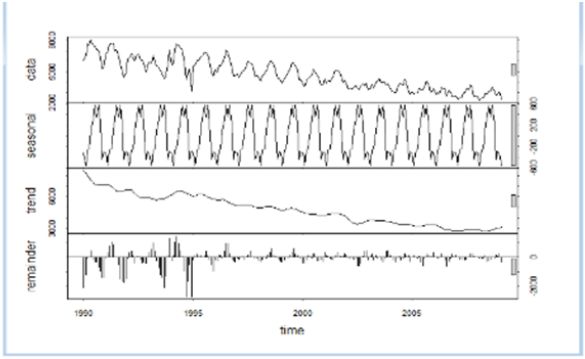
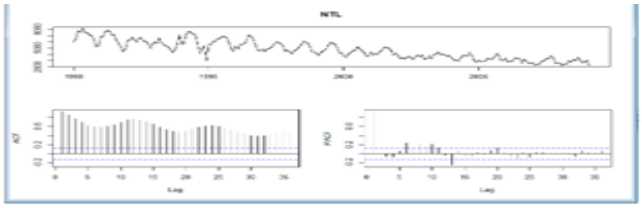


Chart 2. Chart of the decomposition of the NTL series

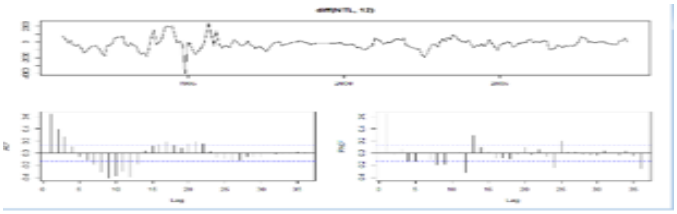
From the decomposition graph, it appears that, in general, the trend is downward, however, the seasonal component is more dominant and the time series has many fluctuations. Generalized Dickey-Fuller test (ADF) The ADF test was implemented, which goes through two stages. In the first stage, the number of lags in each location in the ADF test is determined by the value of the order p of the AR process. it Behe automatically vIA ADF test in R.

```
> ADF_fertility

        Augmented Dickey-Fuller Test

data:  fertility
Dickey-Fuller = -8.2503432, Lag order = 6, p-value = 0.01
alternative hypothesis: stationary
```

Below the resulting ep value Mee SmAll that the level of importance5%, so that, falls under the null hypothesis of non-stationarity AND is confirmed iNTERPRETATION graph. SERIES in the study there is no need for differentiation non-seasonal [6].
It is concluded that the number of non-seasonality is 0.


Graph 3. Correlograms of ACF and PACF for NTL

The ACF correlogram shows that there is a non-stationary seasonal component. It seems, the series has its peak every 12 months and this means that the variable NTL (total fertility rate) must be differentiated 12 times.



Graph 4. Correlograms of ACF and PACF after differentiation for NTL

The first seasonal difference has stationary the series and the correlograms present the presence of autoregressive coefficients and moving average for both seasonal and non-seasonal components. The seasonal order of the moving average is determined by considering only seasonal lags with frequency = 12 and Qmax = 1. Are considered. The PACF plot suggests a non-seasonal maximum trend autoregressive order of 1 ($p_{max} = 1$) and a maximum seasonal autoregressive order ($P_{max} = 2$).

auto.arima(NTL, d = 0, D=1,max.p = 5, max.q = 5, max.P = 2,max.Q =2, max.order = 5, max.D = 1, start.p = 2,start.q = 2, start.P = 1, start.Q = 1, stationary = FALSE,seasonal = TRUE, ic = c(" aicc" , "aic" , "bic"), stepwise = TRUE, trace = TRUE, lambda=T, allowdrift=F)

2. NTL's SARIMA model

2.1. Model Identification

Today there are many procedures or algorithms in different software that enable an efficient and fast pattern identification or recognition [7]. Pattern recognition algorithm The HK algorithm adapts to a random selection test for non-seasonal parameters. From the models identified in R, the best model resulted: *Best model: ARIMA (1,0,2) (1,1,1]* [8]**.**

2.2. Model evaluation

After testing the models with the HK algorithm, the best model was SARIMA (1,0,2) (1,1,1). The next step is to estimate the model parameters using the maximum likelihood method. Using R the results are:

```
Series: NTL
ARIMA(1,0,2)(1,1,1)[12]

Coefficients:
          ar1      ma1     ma2     sar1     sma1
      0.8756034  -0.0245262  -0.1125635  0.0324359  -0.6870882
s.e.  0.0531682   0.0839736   0.0804852  0.0943935   0.0802101

sigma^2 estimated as 209975.3:  log likelihood=-1646.62
AIC=3305.23   AICc=3305.63   BIC=3325.54
```

Above are seasonal and non-seasonal moving average and autoregressive estimates. Thus, it was established that all parameters are significantly different from 0. The model is suitable with respect to the order of the parameters found. By substituting the found values, the final SARIMA model is created:

$$y_t = 0,8756y_{t-1} + 1,0324y_{t-12} - 0,9039y_{t-13} + 0,0324y_{t-24} + 0,02837y_{t-25} + \varepsilon_t - 0,6871\varepsilon_{t-12} - 0,0245\varepsilon_{t-1} + 0,01683\varepsilon_{t-13} - 0,1126\varepsilon_{t-2} + 0,0773\varepsilon_{t-14}$$

2.3. Validation of the model

The SARIMA model fitted so far seems plausible. The seasonal and non-seasonal autoregressive coefficients are significantly different from 0, as are the non-seasonal and seasonal moving average coefficients. The model minimizes the information criteria AIC, BIC and AICc. The next step is to analyze the residuals from the Ljung-Box test.

2.4. Ljung Box test results

```
> Box.test(res_fertility, lag=1, fitdf=1, type="Ljung")

        Box-Ljung test

data:  res_fertility
X-squared = 12.866055, df = 7, p-value = 0.07543996
```

The Ljung-Box test value is greater than the 5% significance level. Therefore, the hypothesis that there is no autocorrelation in the residuals is accepted. Then the stability of the model is checked by means of the inverse of the roots of the characteristic AR and MA polynomials. If all root inverses belong to the complex unit disc, the model is stationary stable.

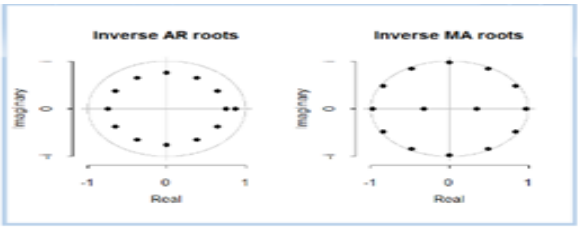


Fig. 1.1. Inversion of Roots

From the figure it is clear that the inverse of the roots of the AR and MA polynomials belong to the complex unit. Therefore, it is concluded that the constructed model is stationary.

2.5. ARIMA forecast

Using the *forecast() function*, ARIMA model predicted future values and confidence intervals for the test observations are constructed. The training data is the first data series, from January 1990 to February 2009, so 230 values [9].

The continuous part of the series, the data from March 2009 to December 2013, i.e. 58 values, was taken as the test set. Since the training set needs more data for model creation than the test set, the ratio between them is 80% to 20%.

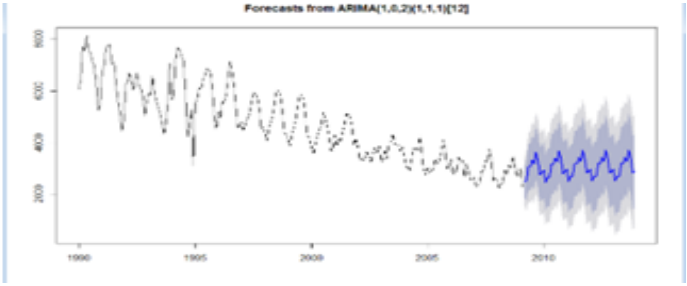
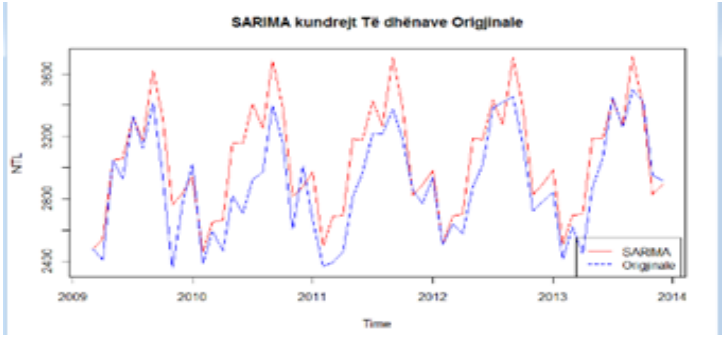


Chart 5. Forecast with NTL's ARIMA model

As can be seen from the graph, the predicted values in relation to the test data follow the trend and seasonality of the series. The confidence interval is also important [10]. It shows that accurate predictions can vary within that range (marked in blue in the figure). Below is the graph of the values predicted by the SARIMA model against the test data of the total fertility rate, in order to compare them.



Graph 6.SARIMA versus NTL time series

The quality of the SARIMA model is shown graphically, where it appears that the fitted values of the model follow the components of the original series to some extent.

2.6. The NAR Model of the Total Birth Rate (NTL)

In the case of modeling this series, the data are divided into two groups: training group and test group [11]. The training data will be used to create the model, the test data to evaluate the created model. The division into training and testing communities is the same as in the ARIMA model. From the HK algorithm that was used to minimize the information criteria, the SARIMA model (1,0,2) (1,1,1) resulted as the best model.

This model is characterized by the presence of the non-seasonal moving average that cannot be taken into account by the NAR (p, k) model. Thus, first the order p = 1 and P = 1 is kept and then the coefficients will be changed to obtain the best NAR model. The function in R is: nnetar (STG, p = 1, P = 1, size = 2, repeats = 20, lambda = TRUE) The 'repeats' option

defines the number of cycles of the neural network training algorithm to find the best architecture. 'Lambda' performs the Box-Coxt transformation and tests if Lambda = TRUE and the 'size' function determines the number of neurons in the hidden layer. The number of neurons in the hidden layer can also be changed, determining the architecture of the network that best fits the model.The function has selected the NAR (4,4,1) model as the best model. Below is the architecture of this neural network.

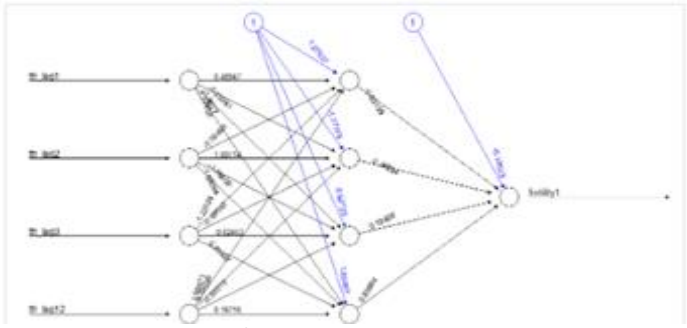
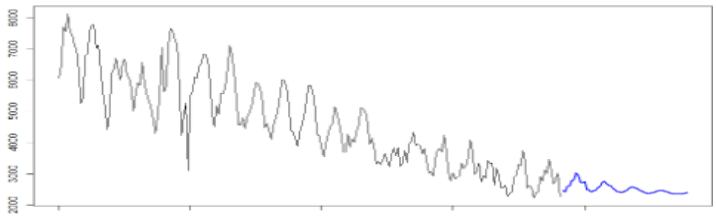


Figure 1.2 Architecture of the neural network NAR time repetition NTL

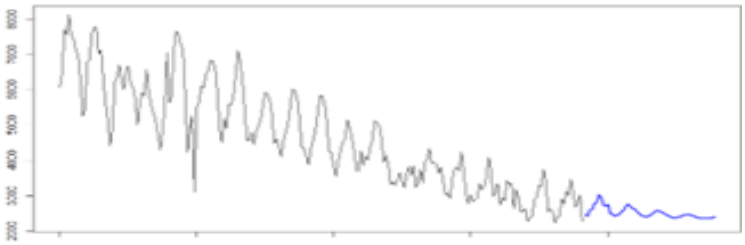
There are four input variables to the network:
tfr_lag1: First time series lag of the total fertility rate
tfr_lag2: Second time series lag of the total fertility rate
tfr_lag3: Third time series lag of the total fertility rate
tfr_lag_12: Seasonal lag of the total fertility rate time series

The network is trained with the R Prop (Resilient Backpropagation) algorithm. It is the same as the *backpropagation algorithm*, but more complex, more accurate and faster in training. This algorithm does not need to determine the learning rate and *moment coefficient. R* It is a heuristic supervised learning algorithm for direct networks, suitable for the type of NAR networks that were used. Since the neural network used is the NAR network (given earlier in the neural networks paragraph) then: the activation function for the hidden layer is the sigmoidal logistic function and the activation function for the output layer is the linear function. The error function is SSE (sum of squared errors) [12]. In the following, it is shown that how the NAR model predicted the 58 values of the test community of the N TL time series.



Graph 7. NeuralNAR network prediction for NTL time series

At the end, the comparison between the values of the NAR model and the data of the original series is presented graphically.



Graph 8. NeuralNAR network versus the original NTL series

The graph above visually represents the inappropriate NAR model for forecasting the NTL time series. The first observations are predicted well, but then the model fails to fit the given observations. The reason lies in the importance of the random component that is not considered in the NAR model. In the SARIMA model the time series contains a non-seasonal component and a seasonal component of the moving average. These coefficients are not allowed in the NAR model because of the way its architecture is built and that is why the prediction is far from the desired one.

2.7. ARIMA-ANN Hybrid Model of the Total Fertility Rate (TFR)

For the construction of the hybrid model, the neuralnet () package was used, among others. Taking into account the proposed hybrid model from the chapter of hybrid explanatory models, this model would start from autoregressive coefficients and moving average coefficients to determine the inputs of the neural network [13]. This process was carried out in ARIMA modeling. Next, the work continues with modeling according to the neural network. The neural network architecture is presented. The network contains three layers, input layer, hidden layer and output layer with neurons as below in the table. As can be seen from the table, there are 10 variables that function as inputs, which are the explanatory variables stored in the SARIMA model, where 5 are time lags and the other 5 variables are the residuals of the SARIMA model [4].

3. Conclusions

IN this work u place analysis of three series time by side THE METHODS QUANTITATIVE ABOUT THE modeled AND then predicted _ variables ECONOMIC THE COURSE THE exchange, rate interest basis _ AND norm total birth rate. The first method that was used was the traditional ARIMA statistical method. It is a method that is easy to apply and predicts satisfactorily in the short term. ARIMA is one of the most widely used methods in time series analysis. This happens thanks to the Box-Jenkins methodology, which is a well-proven method in modeling and extracting the correct parameters of the model. Modeling with ARIMA proves one of the raised hypotheses. The time series under study are SARIMA models.

- Total fertility rate: SARIMA (1,0,2) (1,1,1) [8].

In this study, ARIMA manages to recognize well the seasonal volatility of the series, with somewhat satisfactory modeling and forecasting performances. However, there is still room for improvement in the models.With the hypothesis that the problem lies in the

presence of the non-linear component, modeling using non-linear methods such as ANN was tried. There are several advantages to using ANN for data analysis:

- Once trained, the neural nets produced performance that did not degrade when presented with data.
- The user is not required to set the importance of the variables, as the network itself makes these decisions. Although data and time are limited, the neural network is a very promising method for providing predictive data.

Other positive characteristics differentiate the neural network model from the ARIMA model. One of these characteristics is that the neural network can be updated (learning supplement) without the need for total retraining. To adapt nonlinear structures, these methods perform much better than linear models. Among the neural networks, the nonlinear autoregressive neural network (NARNN) was selected. This is the neural network that was implemented for modeling and forecasting the time series in the study. The stages of building such a network were defined.

The training algorithm was applied until the network model was able to approximate the unknown model for each of the series. In the fertility rate series, the modeling by the NAR network exhibits deviations from the original fertility series data. The models that were built for the time series proved the hypothesis raised regarding the power of the neural network in forecasting. A fact that can negatively affect univariate time series forecasting with a neural network is not specifying the inputs correctly.

However, this, in this study, is solved by using the neural network type NAR, which has the inputs defined by the autoregressive terms obtained from the previous ARIMA modelling. However, series do not only exhibit non-linear behavior, they are combinations of linear and non-linear structures. Therefore, only a linear or non-linear method cannot be used.

The ARIMA model models the linear part and the ANN model continues modeling either for the linear part or for the remaining non-linear part by performing the forecast for the entire series. From the studies and research carried out, this is the first in Albania that uses the hybridization of two techniques, statistical techniques and intelligent techniques for a common goal, the optimization of time series forecasting.

- The hybrid model proposed in this study has advantages over other existing ARIMA-ANN hybrid models, as it does not start from assumptions.
- The hybrid model of the total fertility rate has improved the performance (according to RMSE) by 98.6% by ARIMA and by 99.4% by N AR; (according to MAPE) has improved performance by 98.7% from ARIMA and 99.4% from NAR. As for the comparison of the NAR network with the ARIMA model, the difference was very small, resulting in the NAR network model being slightly more efficient.

Time Series Analysis for Economic Forecasts Using Quantitative Methods and Neural Intelligence still has room for improvement. By changing the number of hidden neurons, the number of layers and the training algorithm, an even better model could be obtained.

Since the focus of this study was the optimization of the prediction by means of the hybrid technique, we did not focus on the optimization of the ANN architecture, but on the optimization of the predictive model by means of linear and nonlinear hybridization. In conclusion, the hybrid ARIMA-ANN model results as the best model in suitability and prediction of the time series in the study, surpassing the performance of the quantitative ARIMA method and the intelligent ANN method. Building and implementing a hybrid ARIMA-ANN method is an effective way to analyze, model, and forecast time series. In such a submodel, the SARIMA model fits the non-stationary linear component and the neural network fits the nonlinearity. This study showed that the hybrid ARIMA-ANN model is a reliable tool for forecasting. Hybrid models (linear and non-linear) are the best models for time series forecasting. Better modeling and forecasting of the series under study was achieved. The efficient ARIMA-ANN model can be used for forecasting by other authors in the interested fields. These results confirm the hypotheses raised at the beginning of the study.

References

[1] C. Granger and A. Anderson, An Introduction to Bilinear Time Series Model. Vandenhoeck and Ruprecht: Guttingen, 1978.

[2] A. Gjonça, A. Aassveand and I. Mencarin, "The highest fertility in Europe - for how long?" The analysis of fertility change in Albania based on individual data", University of Essex, Colchester, UK, 2006.

[3] E. Gjika, "Time series, downscaling, forecasting: studying similarities through series downscaling, methodsax," *Faculty of Natural Sciences. University of Tirana,* pp. 29-33, 2014.

[4] W. Goh, C. Lim and K. Pen, "Predicting Drug Dissolution Profiles with an Ensemble of Boosted Neural Networks : AT ime Series Approach.," *IEEE T Transactions on Neural Networks,* vol. 14, no. 2, pp. 459-463, 2003.

[5] W. W. Guo, M. M. Li, G. Whymark and Z.-X. Li, "Mutual complement between statistical and neural network approaches for rock magnetism data analysis," *Expert Systems with Applications,* vol. 36, no. 6, p. 9678–9682, 2009.

[6] C. Hamzacebi, "Improving artificial neural networks' performance in seasonal time series forecasting," *Information Sciences 178,* pp. 4550-4559, 2008.

[7] G. Grole can and H. Wickham, R for Data Science", O'Reilly, 2016.

[8] F. Hamilton, A. Lloyd and K. Flores, Hybrid modeling and prediction of dynamical systems, University of Virginia, UNITED STATES: https://doi.org/ 10.1371 / journal .pcbi.1005655, 2017.

[9] W. W. Guo and H. Xue, "Crop Yield Forecasting Using Artificialial Neural Networks: A Comparison between Spatial and Temporal Models," *Hindawi Publishing Corporation Mathematical Problems in Engineering,* 2014.

[10] Z. Hajirah imi and M. Khashei, "Improving the performance of financial forecasting using different combination architectures of ARIMA and ANN models," *JIEMS Journal of Industrial Engineering and Management Studies,* vol. 3, no. 2, pp. 17-32, 2016.

[11] S. Cave Ladze, Evaluating methods for time-series forecasting; Applied to energy consumption predictions for Hva leave, Ostfold University, Norway, 2015.

[12] K. Hipel and A. Mc eod, Time Series Modeling of Water Resources and Environmental Systems, Amsterdam, Elsevier, 1994.

[13] F. Huo and A. Poo, "“Nonlinear autoregressive network with exogenous inputs based contour error reduction in CNC machines," *International Journal of Machine Tools and Manufacture,* vol. 67, p. 45–52, 2013.