FORECASTING UNDER APPLYING MACHINE LEARNING AND STATISTICAL MODELS

by

Azhari A. ELHAG[a] and Hanaa ABU-ZINADAH[b*]

[a] Mathematics and Statistics Department, Faculty of Science,
Taif University, Taif, Saudi Arabia
[b] University of Jeddah, College of Science,
Department of Statistics, Jeddah, Saudi Arabia

Original scientific paper
https://doi.org/10.2298/TSCI20S1131E

In a different area of a field of the real life, problem of accurate forecasting has acquired great importance that present the interesting serve which led to the best ways to achieve a goal. So, in this paper, we aimed to compare the accuracy of some statistical models such as Time Series and Deep Learning models, to forecasting the fertility rate in the Kingdom of Saudi Arabia, the data source is the World Health Organization over the period of 1960 to 2019. The performances of models were evaluated by errors measures mean absolute percentage error.

Key word: *time series, deep learning, fertility rate, statistical models, forecasting, error back propagation network , recurrent neural network, long short-term memory, root mean square error*

Introduction

Forecasting is defined as the one of inferential statistics methods, which aims to know what the magnitude of a phenomenon under the study over a period of time, based on data collected and recorded during a previous and successive time period [1, 2]. The importance of forecasting is due to its high role in various life activities, as economic policies [3, 4]. It should be noted that the important role of the prediction process stems from the accuracy of the prediction results [5] that are mainly based on the correct construction of the model generated from these results, where its effectiveness is determined by achieving a set of satisfaction statisticians and pass a series of tests based on the reliability of the relationship [6].

This current study deals with the forecasting of the fertility rate in Kingdom of Saudi Arabia (KSA), as fertility rate is one of the most important demographic factors that effect on the size and structure of the population.

Recently, many of the studies interested with forecasting using statistical and deep learning modeling for many economic, social and natural phenomena in general, and in particular in the field of demographic studies. The population forecasting using ANN is discussed in [7], that developed MLP with GDL to forecast Indian population with high sufficiency accurate. The hybrid model of ANN in [8] from ANN is applied under

* Corresponding author, e-mail: hhabuznadah@uj.edu.sa; a.alhag@tu.edu.sa

ARIMA models to satisfy forecasting model with high accurate model than ANN. Under real data sets, the empirical results show that, the proposed model has present a suitable method satisfies accuracy forecasting achieved by ANN. The results presented by [7, 8] have shown that the neural network have a very high accuracy to predict time series data. For the join approach to time series forecasting proposed and applied by [9]. The two linear and non-linear models (ARIMA and ANN), respectively, are used jointly, with the goal of obtaining various forms for relationship for the time series data. This hybrid model derives its advantage from the unique power of ANN and ARIMA in non-linear and linear modeling. The merging methods can be effective in improving predictive efficiency in the case of complicated problems, done with linear and non-linear correlation structures. The performance of forecasting can be improved with effectively way under combination method. The empirical results with three real data sets clearly show that this hybrid model is able to out-perform each component model used as a single. More advantage of DWT method discussed by [10] which proposed a new method of forecasting by detach a time series data set into non-linear and linear components through DWT. The applications of time series model are fitting for forecast the TFR in Malaysia by [11]. The ARIMA models and AR models are considered and the forecasting performance of these models is evaluated by using post sample forecasting accuracy criterion. It is found that the AR model appeared to be the most appropriate model for forecasting the TFR in Malaysia.

The results given by [12] show that some methods to forecast TFR of India through an approximate Bayes analysis using ARIMA model. The congruent results based on classical pattern are also obtained especially using maximum likelihood estimators [13]. For the problem of forecasting age-specific mortality, the multilevel functional data method has present very efficacious method for two or more people in advanced countries that has very good vital registration systems. To determine the population-specific residual trend and common trend amongst populations. Also, the multilevel functional principal component is used in the analysis of aggregate and population specific data. An R-package fore mortality smooth using P-splines discussed and developed by [14]. Mortality smooth package present two prime functions used to fit data for one or 2-D setting. The P-splines are the method that useful for smoothing mortality trends also this methodology is useful for analyzing the mortality distribution of the age and time. Provides an overview of the design, aims, and principles of Mortality Smooth, as well as strategies for applying it and extending its use. The package mortality smooth considered as a framework for smoothing count data in both two and 1-D settings. The total number of the deaths over the specified age and year interval is supposed to be distributed as Poisson-distribution, and P-splines and generalized linear array models they are employed as a convenient regression methodology. The main topic of the main deep learning techniques presented by [15], and any time series application, analysis and summarized of data. The results illustrate that it's clear that deep learning has a lot to contribute to the field. Also, it has used three time series data for prediction [16]. In this sense the statistical inference based on Fisher information matrix has different applications in quantum physics and information theory [17-19]. The estimation and predication models have various applications to real data and Statistical distributions [20-25]

Many linear and non-linear models have been developed for these time series prediction. The approach in this work is to comprise a robust predictive model for next day ahead prediction. Comparative had been study of traditional artificial neural network: EBPN and deep neural network: RNN and LSTM for financial time series prediction. It had been observed that seep neural network was outperforming the traditional neural network EBPN. The performances of models are measured by error measures: MAPE and RMSE.

Methodology

The ARIMA time series

ARIMA Model Building Steps, fig. 1:

Some notation applied in ARIMA

- with p autocorrelation ordered: $AR(p)$,
- with d difference: $I(d)$, and
- with q moving average ordered: $MA(q)$.

The AR Processing: ARIMA with (P, 0, 0):

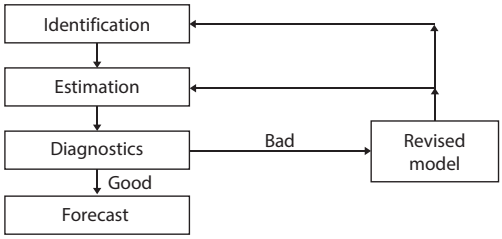

Figure 1. Steps of ARIMA building model

$$z_t = \theta + \phi_1 z_{t-1} + \phi_2 z_{t-2} + \ldots + \phi_p z_{t-p} + \varepsilon_t \tag{1}$$

where ε_t , ϕ.....ϕ_p results in different time series patterns.

The MA Processing: ARIMA with (0, 0, q):

$$z_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q} \tag{2}$$

Integrated process: ARIMA with (0, 1, 0):

$$z_t = z_t + \varepsilon_t \rightarrow z_t - z_{t-} = \varepsilon_t \quad \rightarrow \nabla z_{t=} \tag{3}$$

The ARIMA (p, 0, q):

$$z_t = \phi_1 z_{t-1} + \phi_2 z_{t-2} + \ldots + \phi_p \varepsilon z_{t-p} + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q} \tag{4}$$

Smoothing with single exponential

For forecasting data, this method is more suitable under no seasonal pattern or clear trend [26]. The component form of simple exponential smoothing is given:

$$\text{Forecast equation} \quad \widehat{z_t} + \frac{h}{t} = \ell t \tag{5}$$

$$\text{Smoothing equation} \quad \ell t = \alpha z t + (1 - \alpha)\ell t - 1 \tag{6}$$

where ℓt is the smoothed value of the series at time t.

Neural network model

The ANN are discussed with forecasting methods that are constructed on straightforward mathematical models of the brain [27]. The ANN described as a network of neurons, which are organized in layers. The inputs or predictors are the from the lower layer, then the outputs from upper layer (the forecasts). There may be also amidst layers which contained hidden neurons. Relative to the evolution of the multi-layer concept ANN are being chosen as a tool for performing the prediction [28]. When we add amidst layers with hidden neuron, the neural network will become non-linear see fig. 2.

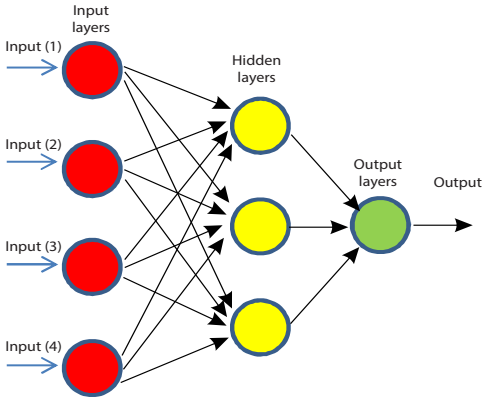

Figure 2. Architecture of neural network

The forecasts will be obtained from a linear combination of the inputs are modified with a non-linear function for example sigmoid:

$$s(z) = \frac{1}{1+\exp(-z)} \tag{7}$$

If the second layer is given as input, which reducing the effects of extreme input values, and prepared the network to be somewhat worthiness to outliers with time series data. Then, we used lagged values of the time as inputs to a neural network, which similer to lagged values in a linear auto regression model which denoted by neural network AR.

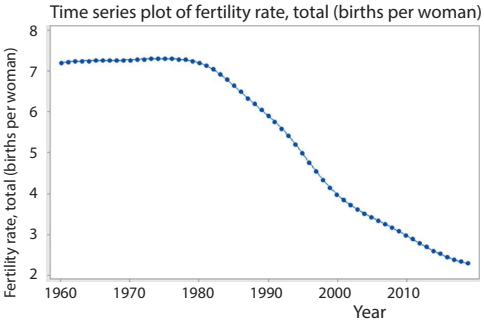

Figure 3. Time series plot of FR

Data analysis

In this paper the modes evaluation is based on the data of FR on sixty years historical demographic time series data using different models. Figure 3 shows the plot time series of fertility rate in the period of 1960-2019 in KSA, as shown the FR declined deeply.

The ARIMA time series model is used for forecasting FR. In this study the best fitting ARIMA model for KSA FR data is identified as ARIMA (0,3,5). The predicted ARIMA model gives a good fit to FR of KSA data and shows good performance therefore, improvements in the ARIMA model is needed to obtain better predictions for FR of KSA data

The ARIMA model: Fertility rate

In tab. 1 we present a modified Box-Pierce (Ljung-Box).

Table 1. Model statistics

Model	Number of predictors	Model Fit statistics			Ljung-Box $Q(18)$		
		Stationary R-squared	MAPE	MAE	Statistics	DF	Sig.
Fertility rate, total (births per woman)-Model_1	1	0.715	0.04	0.002	18.882	13	0.127

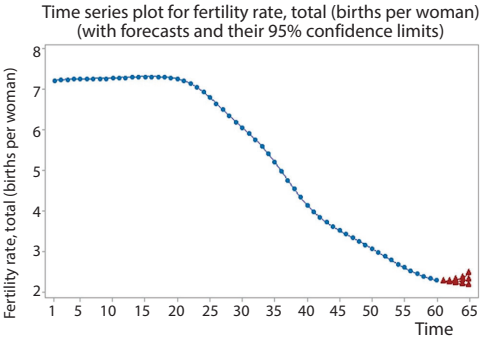

Figure 4. Time Series plot of FR with forecasts from period 60 and their 95% confidence limits single exponential smoothing for fertility rate

As shown from tab. 2 the forecasts from period 60 their 95% confidence limits, fig. 4.

Single exponential smoothing for fertility rate

On the following the plot for smoothing show the fits is closely follow the data, which satisfies the degree of fitting with this model, fig. 5.

From the aforementioned tabs. 3-5, we note the stability of fertility rate.

Multilayer perceptron

Randomly assign cases based on relative numbers of cases has been partitioned in is 75 % training and the testing is 25 % to obtain the model.

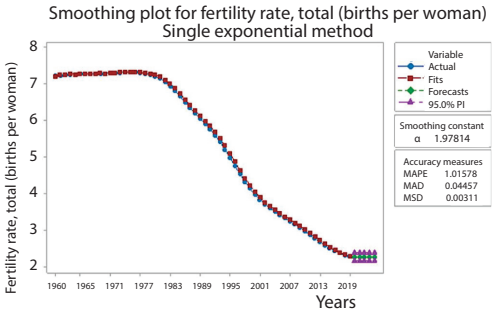

Figure 5. The plot for smoothing of FR with forecasts and their 95% confidence limits

The activation function is hyperbolic tangent and the number of hidden layers, one the activation function for the output is sigmoid. As shown from the model summary, tab. 5. The relative error allmost is zero for both training and testing which is show high accuracy to predicted fertility rate, total (births per woman), fig. 6.

The tab. 6 illustrates the accuracy of each model.

Conclusion

This paper concerned with applying time series model in doing fertility rate forecasting, compared the accuracy of the ARIMA, SES, and MLP. We used the MAPE to measure the accuracy of the model. It is average of the absolute value of percentages error of the forecasting, the lower its value, the more accurate model. The results of this paper show that the MLP is more accurate with MAPE value 2% compared to ARIMA and SES.

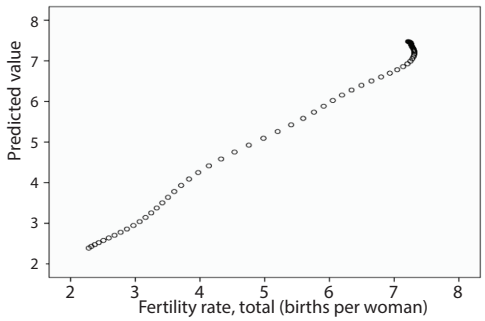

Figure 6. Predicted value for fertility rate

Table 2. The 95% limits

95% limits			
Period	Forecast	Lower	Upper
2020	2.25447	2.24914	2.25980
2021	2.24373	2.22463	2.26283
2022	2.24837	2.20274	2.29400
2023	2.26957	2.18133	2.35781
2024	2.30463	2.15569	2.45357

Table 3. Forecasts

Period	Forecast	Lower	Upper
2020	2.26462	2.15542	2.37382
2021	2.26462	2.15542	2.37382
2022	2.26462	2.15542	2.37382
2023	2.26462	2.15542	2.37382
2024	2.26462	2.15542	2.37382

Table 4. The summary of the case processing

		N	Percent
Valid		60	100.0%
Total		60	
Excluded		0	
Sample	Training	45	75.0%
	Testing	15	25.0%

Table 5. Model summary

Training	Sum (SE)	0.014
	(RE)	0.004
	Stopping (RU)	1 consecutive step(s) with no decrease in errors
	Training Time	0:00:00.02
Testing	Sum (SE)	0.002
	(RE)	0.003

Dependent Variable: Fertility rate, total (births per woman)
a. Error computations are based on the testing sample

Table 6. The accuracy of each model

	MAPE
ARIMA	0. 040
Single exponential smoothing	1.01578
MLP	0.024878

Nomenclature

AR – autoregressive
ARIMA – auto-regressive integrated moving average
DWT – discrete wavelet transform
EBPN – error back propagation network
FR – fertility rate
GDL – generalized delta learning
LSTM – long short-term memory
MA – moving average
MAPE – mean absolute percentage error
RMSE – root mean square error
RNN – recurrent neural network
SES – single exponential smoothing
TFR – total fertility rate

References

[1] Box, G. M. J., *et al.*, *Time Series Analysis: Forecasting and Control*, John Wiley and Sons Inc., Hoboken, New Jersey, USA, 2016
[2] Atul, A., Suganthi, L., Forecasting of Electricity Demand by Hybrid ANN-PSO Models, *Int. J. of Energy Optimization and Engineering (IJEOE)*, *6* (2017), 4, pp. 66-83
[3] Li, R. Y. M., *et al.*, Forecasting the REITs and Stock Indices: Group Method of Data Handling Neural Network Approach, *Pacific Rim Property Research Journal*, *23* (2017), 2, pp. 123-160
[4] Mehdiyev, N., *et al*., Evaluating Forecasting Methods by Considering Different Accuracy Measures, *Procedia Computer Science*, *95* (2016), Dec., pp. 264-271
[5] Coker, F. P., *Understanding the Vital Signs of Your Business*, Ambient Light Publishing, Bellevue, Wash., USA, 2014, pages 30, 39, 42
[6] Lenhard, J., Models and Statistical Inference: The Controversy between Fisher and Neyman–Pearson, The British Journal for the Philosophy of Science, *57* (2006), 1, pp. 69-91
[7] Pandurang T., *et al*., Use of Artificial Neural Networks for Projection of Population of India, *Int. J. of Advanced Engineering and Innovative Technology*, *2* (2015) 1, pp. 2-4
[8] Khashei, M., Bijari, M., An Artificial Neural Network (p, d, q) Model for Time Series Forecasting, *Expert Systems with Applications*, *37* (2010), 1, pp. 479-489
[9] Zhang, G. P., Time Series Forecasting Using a Hybrid ARIMA and Neural Network Model, *Neurocomputing*, *50* (2003), Jan., pp. 159-175
[10] Khandelwal, I., *et al*., Time Series Forecasting using Hybrid ARIMA and ANN Models Based on DWT Decomposition, *Procedia Computer Science*, *48* (2015), Dec., pp. 173-179
[11] Shitan, M., Forecasting the Total Fertility Rate in Malaysia, *Pakistan Journal of Statistics*, *31* (2015), 5, pp. 547-556
[12] Tripathi, P. K., et al., Bayes and Classical Prediction of Total Fertility Rate of India Using Autoregressive Integrated Moving Average Model, *Journal Stat. Appl. Pro*., *7* (2018), 2, pp. 233-244
[13] Shang, H. L., Mortality and Life Expectancy Forecasting for a Group of Populations in Developed Countries: A Multilevel Functional Data Method, *Ann. Appl. Stat., 10* (2016), 3, pp. 1639-1672
[14] Camarda, C.G., MortalitySmooth: An R Package for Smoothing Poisson Counts with 462 P-Splines', *Journal of Statistical Software*, *50* (2012), 1, pp. 1-24
[15] Gamboa, J., Deep Learning for Time-Series Analysis, On-line first, arXiv:1701.01887v1, 2017
[16] Handa, R., *et al*., Financial Time Series Forecasting using Back Propagation Neural Network and Deep Learning Architecture, *Int. J. of Recent Technology and Engineering (IJRTE)*, *8* (2019), 1, pp. 3487-3492
[17] Abdel-Khalek, S., Fisher Information Due to a Phase Noisy Laser under Non-Markovian Environment, *Annals of Physic*, *35*1 (2014), Dec., pp. 952-959
[18] Abdel-Khalek, S., Quantum Fisher Information Flow and Entanglement in Pair Coherent States, *Optical and Quantum Electronics*, *46* (2014), 8, pp. 1055-1064
[19] Abdel-Khalek, S., *et al*., Some Features of Quantum Fisher Information and Entanglement of Two Atoms Based on Atomic State Estimation, *Appl. Math*, *11* (2017), 3, pp. 677-681
[20] Mohie El-Din M. M., *et.a*l., Estimation of the Coefficient of Variation for Lindley Distribution Based on Progressive First Failure Censored Data, *Journal Stat. Appl. Prob*., *8* (2019), July, pp. 83-90
[21] Sabry, M. A., *et al*., Parameter Estimation for the Power Generalized Weibull Distribution Based on One- and Two-Stage Ranked Set Sampling Designs, *Journal Stat. Appl. Prob*., *8* (2019), 2, pp. 113-128
[22] Yusuf, A., Qureshi, Q., A Five Parameter Statistical Distribution with Application Real Data, *Journal Stat. Appl. Prob*., *8* (2019), Mar., pp. 11-26
[23] Kumar, D., *et al*., A New Lifetime Distribution: Some of its Statistical Properties and Application, *Journal Stat. Appl. Prob*., *7* (2018), 3, pp. 413-422

[24] Abonazel, M. R., Different Estimators for Stochastic Parameter Panel Data Models with Serially Correlated Errors, *Journal Stat. Appl. Prob*., *7* (2018), 3, pp. 423-434
[25] Ghazal, M. G. M., Prediction of Exponentiated Family Distributions Observables under Type-II Hybrid Censored Data, *Journal Stat. Appl. Prob*., *7* (2018), 2, pp. 307-319
[26] Kumar, M., A hybrid ARIMA-EGARCH and Artificial Neural Network model in Stock Market Forecasting: Evidence for India and the USA, *Int. J. of Business and Emerging Markets*, *4* (2012), 2, pp. 160-178
[27] Handa, R., *et a*l., Financial Time Series Forecasting using Back Propagation Neural Network and Deep Learning Architecture, *Int. J. of Recent Technology and Engineering (IJRTE)*, *8* (2019), 1, pp. 2277-3878
[28] Jarrah, M., Salim, N., A Recurrent Neural Network and a Discrete Wavelet Transform to Predict the Saudi Stock Price Trends, *(IJACSA) Int. J. of Advanced Computer Science and Applications*, *10* (2019), 4, pp. 155-162

Paper submitted: May 10, 2020
Paper revised: June 1, 2020
Paper accepted: June 10, 2020