The Journal of Animal & Plant Sciences, 28(1): 2018, Page: 290-296
ISSN: 1018-7081

AN INTEGRATED MODEL FOR FORECASTING WHEAT CONSUMPTION IN PAKISTAN: EMPHASIZING SOCIAL AND ECONOMIC FACTORS

N. T. Khan[1], J. Park[2] and Y. B. Kim[1*]

[1]27415 Simulation Laboratory Department of Industrial Engineering, Sungkyunkwan University, Suwon, South Korea;
[2]School of Management and Administration, Yong In University, Yong In, South Korea.
[*]Corresponding author's e-mail: kimyb@skku.edu

ABSTRACT

Wheat is one of the most commonly consumed foods on the Indian subcontinent, especially in Pakistan. It is difficult to find literature regarding wheat consumption forecasting, although some researchers have already focused on wheat production forecasting. Time-series modeling has been adopted with only historical data on wheat, while several influential factors have been neglected. This paper proposes that historical consumption data cannot provide complete information and lacks the power to forecast future values accurately; thus, a simple but powerful model is derived that includes economic and social factors. The proposed model is also compared with conventional approaches, demonstrating its ability to capture highly efficient and accurate forecasts.

Keywords: Forecasting, Social Factors, Economic Factors, Regression.

INTRODUCTION

Wheat is one of the most commonly consumed foods in the Indian subcontinent, especially in Pakistan. Wheat is so widely used that it occupied a 9% share of consumed household items in Pakistan in 2005 (Sher and Ahmad, 2008). In 2015, more than 75% of farmers grew wheat, covering an area of nearly 10 million hectares, or almost half of the total cultivated land, wheat flour constitutes 72% of the daily caloric intake for anaverage person(124 kg per person per year) in Pakistan (Raza, 2015).

Agriculture adds much to the economy of Pakistan, and wheat holds 10% of the total value of the agriculture industry. Wheat held 2.2% of the gross domestic product (GDP) in 2014. Wheat has a key role in the food economy of Pakistan, in terms of both production and consumption (Dorosh and Salam, 2008). Hence, it is important to plan carefully for wheat production to meet the demand. One of the initial and proactive planning steps is forecasting.

Accurate forecasting helps to avoid the tradeoffs of keep-or-sell decisions and the need for urgent imports. Usually, forecasts are based on historical data; however, several factors may affect wheat consumption, one of which is its production in the home country. If the production is less than the demand, imports are carried out; however, imports involve taxes and thus raise the price. Price increases force some people to adopt alternative foods; indeed, changes in wheat prices drive changes in wheat consumption levels (Friedman *et al.*, 2011). Changes in the GDP also encourage changes in food consumption patterns (Gerbens *et al.,* 2010).

Economic factors are worth considering for more accurate demand forecasting. Similarly, certain other factors are usually neglected for forecasting. Social conditions control the social behavior towards food, and rapid increases in the population of a country directly affect the need for food. The ability of a country to feed itself depends on several factors, one of which is population pressure; as the population increases, the food demand rises proportionally (Sadik, 1991).

This paper proposes that historical consumption data cannot provide complete information and lack power to capture accurate forecast values. Thus, a simple but powerful model is derived, which includes two economic factors (the GDP per capita and the average annual price of wheat) and one social factor (the annual population). The results of the proposed model are compared with those of conventional approaches such as auto-regressive-integrated-moving-average (ARIMA) and artificial neural network (ANN) forecasting.

It is difficult to find literature regarding wheat consumption forecasting; however, some researchers have already focused on wheat production forecasting. It will be helpful to review previous studies to understand the conventional approaches to wheat production and the forecasting methods applied. Iqbal *et al.* (2005) employed the ARIMA model to forecast wheat production and the land required to grow it, using production data from the previous 30 years to forecast for the next 10 years. Sher and Ahmad (2008) estimated parameters with the Cobb-Douglas production function for wheat using input values from available resources, and also implemented the ARIMA model to forecast wheat production in Pakistan. For the forecasted period (2007-2015), an average growth of 1.6% was predicted, which was slightly lower than the

expected population growth rate. The authors concluded that the labor force and tractors are factors influencing wheat production. Amin *et al.*(2014) trained a large data set ranging from 1902 to 2005, and compared several models involving different methodologies. The authors ultimately concluded that the ARIMA (1, 2, 2) model was best fit for long-term forecasting of 50 years.

A relatively different approach to time-series modeling is the use of a Moderate Resolution Imaging Spectroradiometer (MODIS). Dempewolf *et al.* (2013) applied satellite-based methodology for wheat production forecasting. This concept is based upon early-season production forecasting, where the wheat land for cultivation is derived from a MODIS satellite-image time series, while the yield is estimated from previous yields against a MODIS-Normalized Difference Vegetation Index (MODIS-NDVI) by means of regression analysis. This novel approach is complex, and it is not feasible for everyone to collect data from MODIS, as it may increase the cost of forecasting. Previously, Becker *et al*. (2010) also used MODIS data to forecast winter yields in Kansas and Ukraine, taking a similar approach with a regression-based model. This method seemed simpler, as it had fewer data requirements.

The relevant studies indicate that there are no such studies on wheat consumption forecasting; however, it is highly important to forecast consumption in order to plan production. Previously, either time-series modeling has been adopted with only historical data about wheat while several influential factors have been neglected, or complex methods have been used.

MATERIALS AND METHODS

As soon as complexity increases, the probability of forecast errors also increases; thus, clients prefer accuracy and only accept forecasts with simple, evidence-based procedures (Green and Armstrong*.,* 2015). It is helpful to consider the importance of simplicity during the model development and selection phase. Model selection is one of the most important steps of any statistical analysis (Kadane*&* Lazar, 2004). A simple methodology is chosen for the simple yet dominant model, as shown in Fig. 1.

Fig. 1. Framework for wheat forecasting

Factors: The first economic factor is the GDP per capita, the purchasing power of the common man. It is calculated by dividing the annual GDP by the annual population of the country. The second factor is the average annual price of wheat. The price of wheat may change every day, which is why the average annual price is used. The third influential factor is a social factor, annual population.

Data Analysis: Historical data are gathered from openly available Internet sources. Data are visualized for the analysis so that the model can be selected. Fig. 2 (a) depicts the scatter plot of wheat consumption in metric tons (MT) at year *t* vs. wheat consumption (MT) at year *t-1*.The plot exhibits autocorrelation with an R^2 value of 0.9922, indicating that the simplest model, order-one auto-regression (AR (1)), can be used. Similarly, the scatter plot of wheat consumption at year *t* vs. annual population (in millions) at year *t-1* is depicted in Fig. 2 (b). The population of the previous year has a linear relationship with the wheat consumption of the current year; hence, the annual population can be used in linear modeling. Unlike annual population, both social factors (GDP per capita and annual average price of wheat) exhibit somewhat linear behavior when plotted against wheat consumption, but the R^2 values do not encourage their use in the linear model. Thus, data preprocessing is necessary.

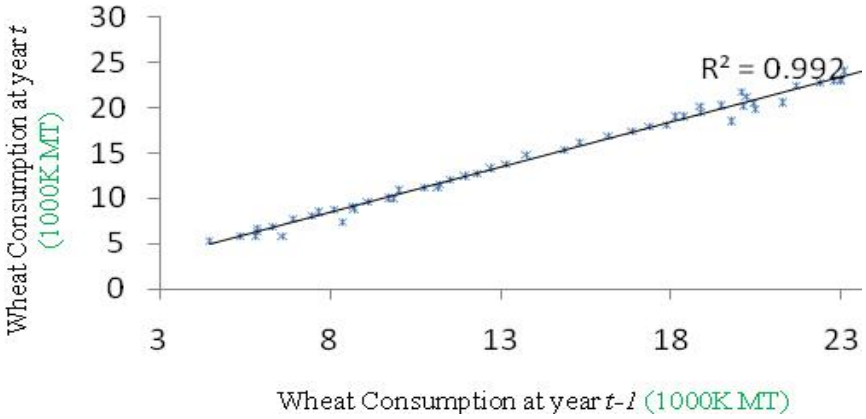

Fig. 2.(a) Wheat consumption at *t* vs. *t-1*

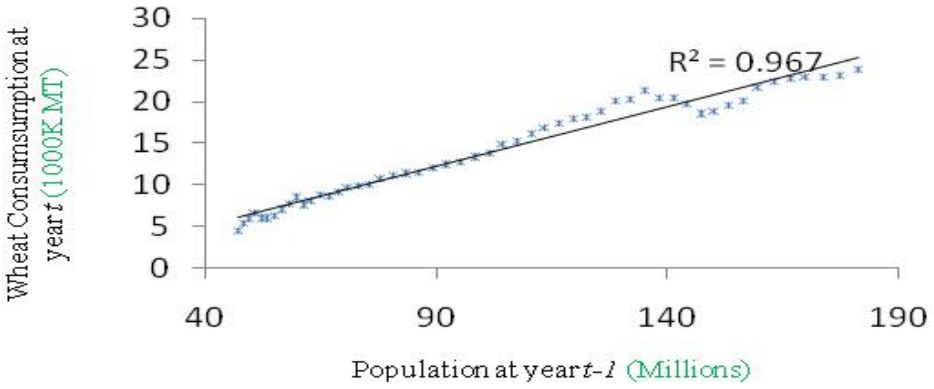

Fig. 2. (b) Wheat consumption at *t* vs. population at *t-1*

Data Preprocessing: Most of the time, raw data contain noise or do not provide exact information. Hence, preprocessing is required to handle such data. Data pre-processing is usually neglected, but is a very important step. Some of the data preprocessing techniques are data de-seasonalization, standardization, cleaning, integration, transformation and reduction. The selection of a data preprocessing technique purely depends on the characteristics of the input data; however, sometimes it is so critical to select the best form of data preprocessing that it is done by trial and error. Data transformation is usually the easiest technique, because techniques for best fit are already available (for example, Box-Cox techniques). Since both economic factors exhibit the same pattern, we apply natural logarithm transformation to both factors.

Fig 3(a) displays the plot of wheat consumption in the current year vs. GDP per capita (in current US dollars) in the previous year, and Fig 3(b) displays the plot of wheat consumption in the current year vs. the average annual price of wheat in Rupees per Mon (RS/Mon) in the previous year. Fig 4(a) and Fig 4(b) represent the plots after natural log transform. The R^2 values are greatly improved, indicating that both economic factors are usable for linear modeling. The R^2 value changes from 0.78 to 0.91 in the case of GDP per capita, and from 0.59 to 0.95 in the case of price.

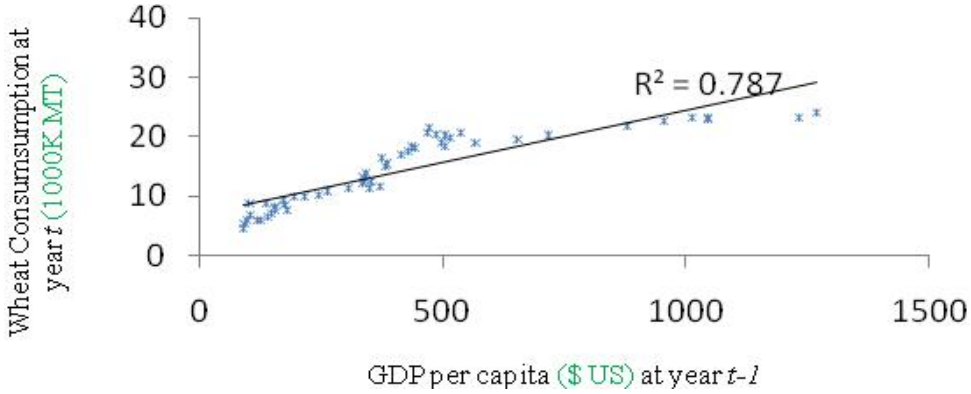

Fig. 3. (a) Wheat consumption at *t* vs. GDP per capita at *t-1*

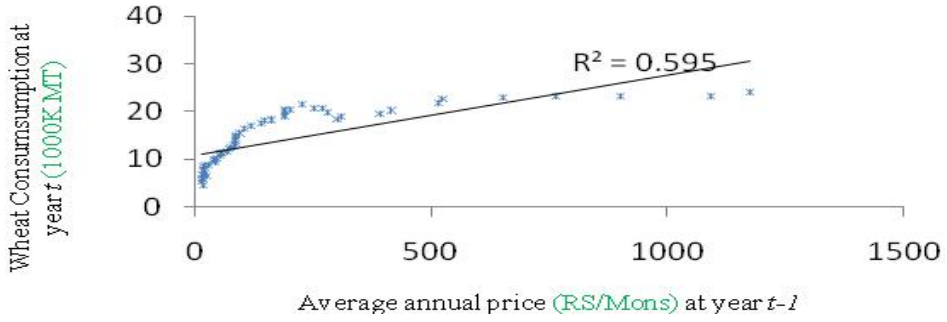

Fig. 3. (b) Wheat consumption at *t* vs. annual price at *t-1*

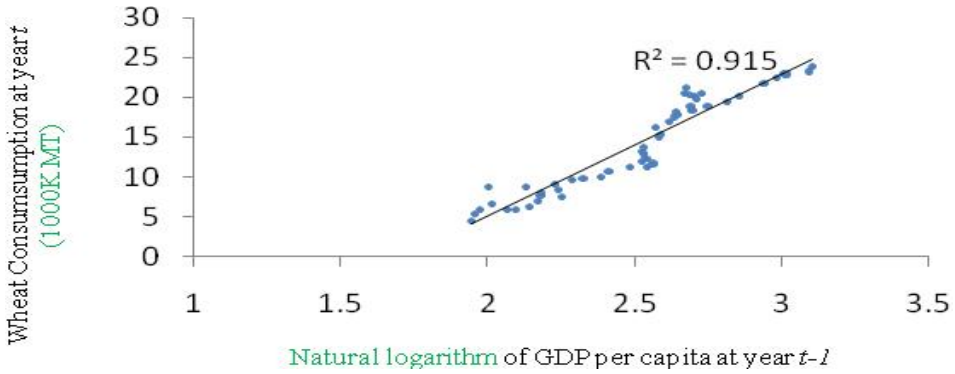

Fig. 4. (a) Wheat consumption at *t* vs. natural logarithm of GDP per capita at *t-1*

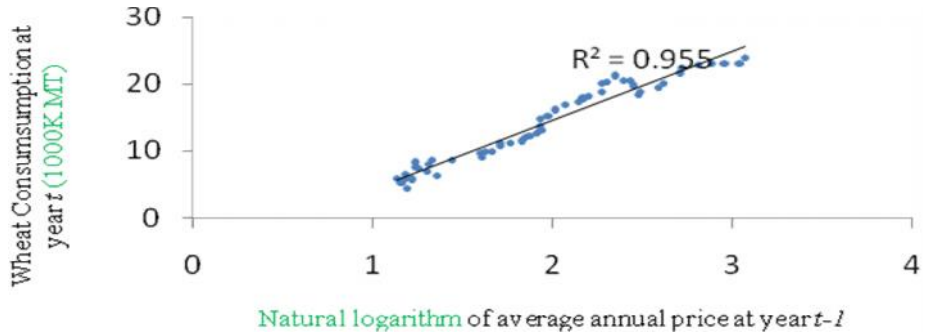

Fig. 4.(b) Wheat consumption at *t* vs. natural logarithm of annual price at *t-1*

RESULTS AND DISCUSSION

First, notations are defined to construct modified linear regression models, as follows:

Z_t: wheat consumption at year t

G_t: GDP per capita at year t

P_t: annual population at year t

R_t: average annual price of wheat at year t

Using the above notations for wheat consumption and social and economic factors, we construct various models by regressing influential factors mixed with AR (1). For instance, if all the factors are input variables, then we obtain the modified regression model including AR (1) as follows:

$$Z_{t+1} = \alpha + \beta_0 Z_t + \beta_1 P_t + \beta_2 \ln(R_t) + \beta_3 \ln(G_t) + \varepsilon_t \quad (1)$$

In the above model, all of the aforementioned factors are included along with historical wheat consumption, where ε_t represents unknown error. In contrast, the AR (1) presented in (2) does not account for any other factor, and is based solely on historical wheat consumption. Thus, AR (1) is tested first, and then each factor is included. Then, combinations of factors are included, so that finally eight different models including AR (1) are tested. Experiments are performed in two different ways for each model: the cumulative method and the non-cumulative method.

$$Z_{t+1} = \alpha + \beta_0 Z_t + \varepsilon_t \quad (2)$$

The available data for wheat and all three factors consist of 53 data points from 1961 to 2013. The data are divided into a training set and a test set. The first 30 data points are considered training data, and the results are tested on the next 23 points. In the cumulative experiments, the number of data points in the training set increases after every prediction; the predicted value is added to the training set, such that the second prediction training set consists of 31 points, and so on. In contrast, in non-cumulative experiments, only the previous 30 points are considered for every prediction.

After the training and test sets and experimental methods are established, it is important to select performance measures. This selection depends highly upon the preferences of the analyst. For this experiment, three different performance measures are selected. One of the most commonly used performance measures is the mean absolute percentage error (MAPE), which reflects the overall forecasting performance. The maximum absolute percentage error (Max.APE) is also used, which indicates the highest error of a forecast. The third performance measure used for the experiment is the 90 percentile, which is adopted if the analyst is required to know about the distribution, and can also be considered as the variance. The results from all eight models are presented in Table 1. The results demonstrate that almost all the models incorporating influential factors are better than the AR (1) model, in which only historical wheat data are used. Table 2 is a more summarized form of the results, indicating that the best value for the MAPE is achieved when AR(1) is mixed with the average annual

price, while the best value for the Max.APE is achieved when AR (1) is combined with both economic factors. The model with population and GDP mixed with AR (1) yields the lowest value for the 90 percentile of APEs among all the models.

Table 1. Experimental Results.

Input Variables				Method	MAPE (%)	MAX.APE (%)	90 Percentile of APEs
Z_t	P_t	R_t	G_t				
o				Cumulative	2.45	9.53	6.01
				Non-Cumulative	2.49	9.41	6.29
o	o			Cumulative	2.84	12.77	5.46
				Non-Cumulative	2.95	10.75	6.11
o		o		Cumulative	2.33	9.97	5.62
				Non-Cumulative	2.56	9.46	6.25
o			o	Cumulative	2.41	9.31	5.72
				Non-Cumulative	2.57	9.32	5.64
o	o	o		Cumulative	2.89	12.65	5.52
				Non-Cumulative	2.99	10.91	6.26
o		o	o	Cumulative	2.41	9.94	5.75
				Non-Cumulative	3.02	8.18	6.95
o	o		o	Cumulative	2.80	12.55	5.35
				Non-Cumulative	3.24	10.24	6.79
o	o	o	o	Cumulative	2.81	12.56	5.39
				Non-Cumulative	3.18	9.65	7.32

Table 2. Summary of best values.

Performance Indicator	Results with Wheat dataOnly	Best Results	Factors Providing Best Results
MAPE (%)	2.45	2.33	Price
Max.APE(%)	9.41	8.18	Price and GDP
90 Percentile	6.01	5.35	Population and GDP

Selection of the cumulative or non-cumulative method is important, as reflected by the results. The method adopted and the results obtained may differ for different data and models. The results are compared in terms of the method in Fig 5. After comparing all eight models in terms of the cumulative and non-cumulative methods, we conclude that the cumulative method performs well overall; however, the results do not encourage neglecting the non-cumulative method, as Fig 5 depicts that this method also performs well in calculating the Max.APE.

The selection of a performance measure and the selection of a method of evaluating historical data are interrelated. The MAPE, Max.APE and 90 percentile are calculated eight times for eight different models. The cumulative method predicts more accurately all eight times for the MAPE, one time for the Max.APE and eight times for the 90 percentile. Overall, for 24 times, the cumulative method provides better results 17 times, or about 70% of the time.

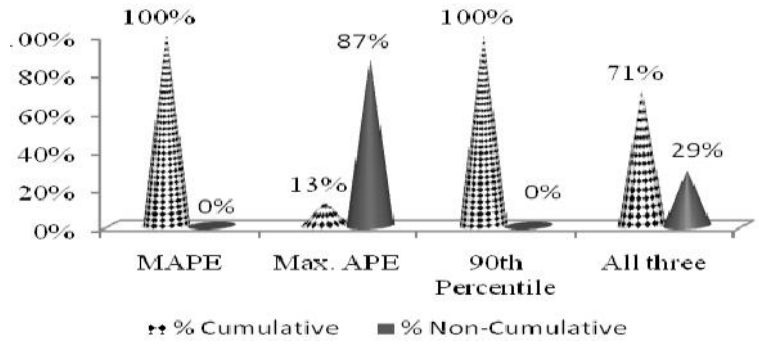

Fig. 5. Comparison between the cumulative and non-cumulative approaches

From the experiment and the detailed analysis expressed in the figures and tables, it is evident that for three different performance measures, three different models perform best. Concerning the choice of the best model among these three models, the model with both economic factors is recommended. This model provides the best value for the Max.APE and the second-best value for the MAPE, and also improves the 90 percentile considerably.

Comparison: Reviewing the related literature, we find that the method conventionally adopted for forecasting in the wheat field is uni-variate time-series forecasting, and ARIMA models are typically used for this purpose. ARIMA models are also used in various other fields because they capture patterns powerfully. Our model is compared with the ARIMA model. An *auto.arima* function in *R software* is used to develop the best fit for the data. The function selects the ARIMA (0, 1, 0) model for the best fit.

One of the most recent forecasting approaches is the ANN, and many studies are being conducted to improve ANNs, as well. A general neural network model is applied to data in *R* software, which selects the NNAR (1,1) as the best fit. Fig 6 depicts the comparison of the ARIMA, the ANN and the proposed regression model mixed with AR (1) for the test set ranging from 1992 to 2014. A graphical presentation of the comparison demonstrates that the proposed model captures the ups and downs of the time series more effectively and efficiently than the other models.

The MAPE values of the three compared methods are summarized in Table 3. The table demonstrates that the proposed model is very strong in terms of the MAPE performance measure. More than half of the predictions have a MAPE under 2%, and more than 80% have a MAPE under 5%.

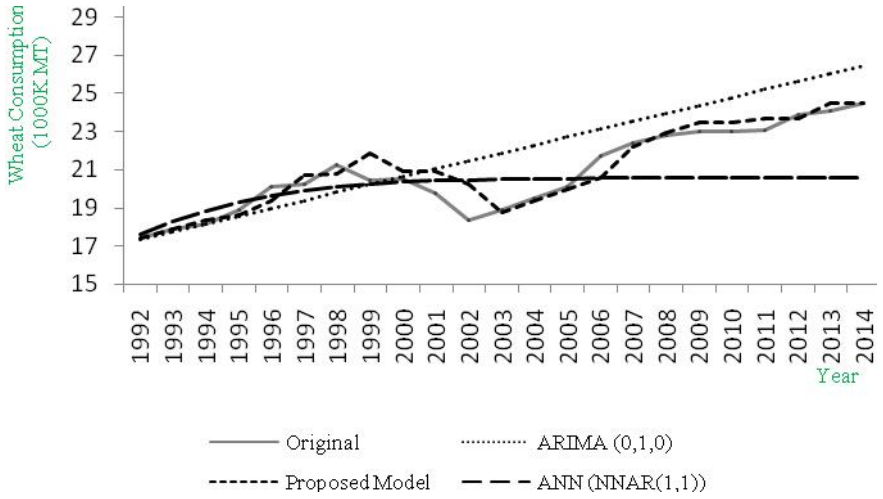

Fig. 6. Comparison of the ARIMA, ANN and Proposed Model

Table 3. Performance of the three methods in terms of MAPE.

Method	MAPE (%)	MAPE under 1%	MAPE under 2%	MAPE under 5%
ARIMA	6.5%	17.39%	26.09%	30.43%
Proposed Model	2.3%	8.70%	17.39%	82.61%
ANN	6.5%	39.13%	52.17%	43.48%

Conclusion: The model of historical wheat consumption and average annual price data provides the most efficient forecast values among all of the models in terms of the MAPE, and is recommended especially for short-term forecasting. For long-term forecasting, the model with historical wheat consumption data and both economic factors seems to be more reliable, as it has the best value for Max.APE. To make this model useful, it is important to have future values for all influential factors, but this is not feasible. Even when there are accurately forecasted future values for all social and economic factors, the chance of error may increase because there are no real values. Thus, it is better to recommend either one or no other factors to be considered, and to use previous wheat consumption data only for long-term forecasting. Comparison of the MAPE values of AR (1), ARIMA and ANN confirms that AR (1) is the best choice among the uni-variate models for long-term forecasting. Thus, based

on the choice of analyst, if a well-distributed forecast is required, the model including historical wheat consumption, population and GDP per capita could be recommended.

The most suitable model with respect to all three performance indicators is the model in which AR (1) is combined with regressing both economic factors (price and GDP per capita). If the analyst wants to consider all three performance measures, then the model with both economic factors can be selected, because this model improves all three indicators relative to the original AR(1) model of historical wheat consumption, as well as ARIMA and ANN. The proposed model yields the best value for the Max.APE, the second-best value for the MAPE and a quite improved value for the 90 percentile.

Although the economic factors have stronger effects on wheat consumption than the social factor, the social factor may also be useful, depending on the method adopted. When these important aspects are considered, accuracy can be improved without the need for complex methods. Future work could produce more powerful forecasts by including further economic and social factors, such as inflation and exchange rates, taxes, the quality of wheat production and imports / exports. Another future aspect of this study is modeling the ARIMA with regressors, where external factors are regressed with ARIMA. The authors are working on training an ANN including external factors like temperature, rainfall and flooding.

REFERENCES

Amin, M., M. Amanullah, and A. Akbar (2014). Time series modeling for forecasting wheat production of Pakistan. The J. Anim. Plant Sci. 24(5): 1444-1451.

Becker-Reshef, I., E. Vermote, M. Lindeman, and C. Justice (2010). A generalized regression-based model for forecasting winter wheat yields in Kansas and Ukraine using MODIS data. Remote Sens. Environ. 114(6): 1312-1323.

Dempewolf, J., B. Adusei, I. Becker-Reshef, B. Barker, P. Potapov, M. Hansen, and C. Justice (2013). Proc. IEEE Int. Geosci. Remote Sens. Symp., 3239-3242.

Dorosh, P., and A. Salam (2008). Wheat markets and price stabilisation in Pakistan: An analysis of policy options. Pakistan. Dev. Rev. 47(1): 71-87.

Friedman, J., S.Y. Hong, and X. Hou (2011). The impact of the food price crisis on consumption and caloric availability in Pakistan: evidence from repeated cross-sectional and panel data, World Bank, Technical report, Washington, DC.

Gerbens-Leenes, P.W., S. Nonhebel, and M.S. Krol (2010). Food consumption patterns and economic growth. Increasing affluence and the use of natural resources. Appetite. 55(3): 597-608.

Green, K. C., and J.S. Armstrong (2015). Simple versus complex forecasting: The evidence. J. Bus. Res. 68(8): 1678-1685.

Iqbal, N., K. Bakhsh, A. Maqbool, and A.S. Ahmad (2005). Use of the ARIMA model for forecasting wheat area and production in Pakistan. J. Agric. Soc. Sci. 1(2): 120-122.

Kadane, J. B., and N.A. Lazar (2004).Methods and criteria for model selection.J. Am. Stat.Assoc. 99(465): 279-290

Raza, A. (2015). Pakistan: Grain and Feed Annual Report, USDA Foreign Agricultural Service, Islamabad.

Sadik, N. (1991). Population growth and the food crisis. Food Nutr. Agri. 1(1): 3-6.

Sher, F., and E. Ahmad (2008). Forecasting wheat production in Pakistan. Lahore J. Econ. 3(1): 57-85.