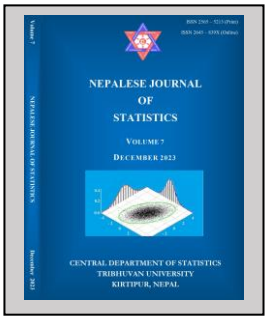
Modeling and Forecasting of Spinach Production in Bangladesh

Keya Rani Das[1], Mashrat Jahan[2]*, Linnet Riya Barman[3] and Preetilata Burman[4]

Submitted: 11 May 2023; Accepted: 30 September 2023
Published online: 26 December 2023
DOI: https://doi.org/10.3126/njs.v7i1.61053



ABSTRACT

Background: In Bangladesh, spinach (*Spinachia oleracea L.*) is most frequently known as "Palong shak" or Bengal Spinach. Spinach is one of the most prominent vegetable crops grown around the earth. It is a quick-growing annual plant that only lives for one year. Although spinach can be cultivated at any time throughout the year, Bangladesh's production of spinach on a per-unit basis is very low in comparison to that of other advanced nations.

Objective: This study explores the impacts of area, price, and weather parameter adaptability on the local production of spinach in Bangladesh employing annual data from time series covering 60 years.

Materials and Methods: This investigation seeks to comprehend the linear relationship between spinach production, commercial price, geographic location, and meteorological characteristics in Bangladesh by applying stepwise regression analysis to find the most suitable model of Spinach production. In order to evaluate how well the multiple linear regression model fits the data, several diagnostic charts were constructed using the R programming language. To forecast the Spinach production, an autoregressive integrated moving average (ARIMA) model was applied.

Results: According to the findings of the study, there was a positive correlation between spinach output and total harvest area under spinach cultivation. There is weak positive correlation between production and annual mean temperature and a weak negative correlation exists between production and annual mean rainfall. Also, production is perfectly positively correlated with price. Regression analysis revealed that the variables sales price and annual mean rainfall are the best predictors of spinach output. Every unit increase in the price of spinach is increases an additional 5.98 times the average of spinach being produced. An increase of 1 millimeter in annual mean rainfall (AMR) decreases 0.001 times spinach production while other predictors remain fixed. Also, ARIMA (1, 1, 0) with drift model forecasts for next fifteen years that the Spinach production is expected to increase considerably by 75000 tons.

Conclusion: Finally, the study found that farmers' perspective on spinach cultivation could potentially experience a positive upward shift which is expected to last for the next fifteen years.

Keywords: ARIMA, predicting, price, rainfall, R-studio, spinach, stepwise regression.

Address correspondence to the author: Department of Statistics, Bangabandhu Sheikh Mujibur Rahman Agricultural University, Gazipur-1706, Bangladesh.
Email: keyadas57@bsmrau.edu.bd[1]; Department of Agricultural Economics, Bangabandhu Sheikh Mujibur Rahman Agricultural University Gazipur-1706, Bangladesh.
Email: mjahan.aec@bsmrau.edu.bd [2*] (Corresponding author email); Faculty of Agricultural Economics and Rural Development, Bangabandhu Sheikh Mujibur Rahman Agricultural University, Gazipur-1706, Bangladesh[3, 4]. Email: linnetriya@gmail.com[3]; Email: preetypbs@gmail.com[4]

INTRODUCTION

Spinach, scientifically known as *Spinacia oleracea L.*, is a yearly dioecious crop that can reach a height of up to 30 centimeters. It is a member of the Chenopodiaceae family and is regarded as an extraordinary green vegetable on grounds of both flavor and nutritional value. The essential amino acids, iron, vitamin A, and folic acid are the main chemical components of spinach. Iron, vitamin A, and folic acid requirements are met daily by consuming raw spinach. Nearly all vitamins are found in spinach leaves, which are also the richest source of beta-carotene and lutein. Carotene is the source of vitamin A. They are one of the finest sources of iron and include nearly all the minerals (Tewani et al., 2016). Around 23,000 acres are used for spinach farming in Bangladesh right now, yielding roughly 55,000 tons and an average of 2,431 kg per acre. Since spinach extract is rich in flavonoids, phenolics, carotenoids and vitamin C (Bergman et al., 2001; Babu et al., 2018; Naznin et al., 2019) it is also acknowledging highly powerful antioxidant activity with a low IC 50 (less than 50 g/mL), which comes out to around 29.67 g/mL (Molyneux, 2004). The growing season for spinach is consistent all year. Spinach is often harvested in the fall months of October and November. At the time when there is a scarcity of veggies because most of the vegetables are still in their early growing stages, the use of spinach helps alleviate this problem. Because of its high nutritional content, it is sometimes referred to as "Life food." While spinach is very nutritious, per unit output in Bangladesh is far lower than in more industrialized nations (Nath et al., 2020). As spinach is considered important vegetables a limited number of researches have already been conducted in Bangladesh on the subject of spinach farming from a variety of perspectives. A semi-log model was used to examine the pattern and rate of increase in acreage, volume, and yield of green spinach in Bangladesh from 1986 to 2016 (Sharmin et al., 2018). In order to figure out the returns of spinach when intercropped with red amaranth and brinjal, pooled data was numerically reviewed, recorded using the LSD test which captured the yield & yield contributing features (Ali et al., 2022). The

analysis of variance technique was used in order to investigate the impact that planting time had on the production and freshness of spinach (Pathania et al., 2022). There is also some research done concerned with environmental issues. There is a possibility that even an increase in precipitation may have a beneficial effect on expansion **(**Grossiord et al, 2020**)**; however, there is also a possibility that intense precipitation could have a detrimental effect on growth and, therefore, yield. Germination of spinach seeds occurs most successfully between 15 and 20 degrees Celsius. Beyond 20 degrees Celsius, germination percentages begin to decline, with a precipitous fall occurring around 35 degrees Celsius (Chitwood et al., 2016)**.** The elevated CO_2 treatment could enhance growth and nutritional value of spinach, and further contribute to CO_2 reduction (Seo et al., 2017). But all of these environmental related articles focused on particular abiotic stress while this default research dedicated on the combination of several environmental factors effect on spinach production. It is difficult to locate any study that prepares a regression model of producing spinach and forecasts its output in Bangladesh, despite the fact that several studies have already been conducted on the cultivation of spinach in Bangladesh. To conduct this analysis on time series data the Stepwise regression was used in this study to draw the best suitable production model and ARIMA were used in order to accomplish the study's primary objective of predicting and accurately forecasting spinach output in Bangladesh.

METHODOLOGY

This study examines the linear relationship among production, commercial price, geographic location, and meteorological characteristics production of spinach in Bangladesh using 60 years of yearly time series data from 1961 to 2020. The data was collected from two secondary sources-FAOSTAT (https://www.fao.org/faostat/en/#data) and Climate Knowledge Portal (climateknowledgeportal.worldbank.org). The following factors have been added to determine the best possible regression line: production in tons, harvest area in hectares, yearly mean rainfall in millimeters, and annual mean temperature in degrees Celsius. To find the best suited model stepwise regression has been applied. Several diagnostic charts were constructed using the R programming language (R v 4.2.1) in order to evaluate how well the multiple linear regression model (MLRM) fits the data. An autoregressive integrated moving average (ARIMA) model was applied to forecast the spinach production.

Model analysis

Several linear regression models have been used in Bangladesh in order to demonstrate the linear connection that exists amongst the production of spinach, its commercial price, its geographic location, and the climatic conditions that affect Bangladesh. The MLRM is an extension of the basic linear regression model that covers data with many predictor variables but only one result. It was developed after the simple linear regression model. Throughout the last 60 years,

Bangladesh's spinach production has been analyzed using a regression-based production model, which allowed researchers to determine how the elements of yield, price, area, and climatic stress affected the increased efficiency of spinach cultivation. The following statistical relationship is formalized between both the single continuous production output and the predictor variables X_k (k = 1, 2,..., p-1)

$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_{P-1} X_{P-1} + \varepsilon_i$ ..... (1)

where $\varepsilon_i \sim N(0, \sigma^2)$, $Y_i \sim N(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_{P-1} X_{P-1}, \sigma^2)$ and σ^2 stands for variance. Evaluation of Pearson's correlation coefficient allows one to assess the degree to which certain factors, such as spinach local production in relation to price, area, yield, and environmental conditions, are correlated with one another. The formula used is as follows: $r_{xy} = \frac{cov(xy)}{\sqrt{var(x) \times var(y)}}$

As it considers the continuous stream of each and every occurrence, the linear regression model requires assumptions about the information it represents in order to remain acceptable **(**Casson et al, 2014**).** There are four presumptions that drive the research. The first step in avoiding erroneous deductions and logical conclusions is to ensure that the normalcy assumption holds. Using quantitative techniques on normality, such as skewness, kurtosis, normal P-P plots, normal Q-Q plots, and the Kolmogorov-Smirnov test, we will determine whether or not this assumption holds true (Das and Imon, 2016). The variables influencing spinach production are most likely to fall into a data is normally distributed, which is a prerequisite for using regression. Correlations and significance tests may be impacted by non-normal data characteristics (highly skewed or kurtosis, or variables with several outliers) (Osbourne and Waters, 2002).

In the second step of the process, the linearity assumption is used to determine whether or not the effects of a number of variables, whether they have been transformed or not, are added together to produce a model with residuals that are frequently and completely dispersed, as well as distributed inconsistently. In the third step, the existence of heteroscedasticity is investigated to determine whether or not the presumption of homoscedasticity was violated in any way. When the magnitude of the residual variance; varies across the values of an independent variable, we say that there is heteroscedasticity in the data. In the fourth step, the outliers, or values that are significantly different from the rest of the points in the data, are identified. Outliers provide a challenge for a wide variety of statistical studies because they have the potential to lead tests to either overlook key discoveries or to skew actual results (Frost, 2019). If a data point has a very high predictor input value, then that point is regarded to be a High Leverage Point. In the event of outliers in the data, the amount of possible damage that may be caused by high-leverage points is increased. The accuracy of the line that represents the least squares solution will suffer significantly if a high-leverage point is also an anomaly. In the subsequent stage, VIF values are employed to zero in on the elements that are causing issues with multicollinearity. Along with VIF values, also applied

tolerance values, eigenvalues and condition index to check the multicollinearity problem. Multicollinearity problem addresses in Hasan at el., 2016. When this has been accomplished, the problematic variables are removed from the regression.

Various analysis methods utilized in this study can be found in existing literature pertaining to the prediction of production for various crops. Stepwise regression is used in order to optimize the estimating power via the utilization of the minimum number of independent variables that is practically possible. An application of stepwise regression is presented in Hasan at el. (2016). Either the forward selection approach or the backward elimination method may be used to carry out the stepwise regression. In this investigation, the elimination of variables that lacked significance was accomplished by the application of the forward selection approach. Singh et al. (2014) formulated prediction equations for rice and wheat yield by utilizing weather and yield data spanning eighteen years (1991-2008) from nine districts in Eastern Uttar Pradesh, India. The stepwise regression procedure was employed to determine the most suitable regression equations from a set of independent variables. This study employs stepwise regression to identify the most appropriate model for predicting spinach production, while minimizing the number of independent variables to a practical minimum.

In the last stage, this time series data was forecasted using auto.arima() function under 'forecast' packages in R Studio. The Autoregressive integrated moving average (ARIMA) model building has the advantages of being adaptable, relatively accurate, and scalable to the study of a large number of time series. The model's consistency and stationary test led us to select annual estimates of spinach production from 1961 to 2020 for this analysis. Model's performance was checked using the Akaike information criterion (AIC), Bayesian Information criterion (BIC), maximum likelihood estimation (MLE), standard error (SE), root mean squared error (RMSE), mean absolute error (MAE), mean percentage error (MPE), mean absolute percentage error (MAPE), mean absolute scaled error (MASE), and the first-order autocorrelation coefficient (ACF1). An ARIMA model is labeled as an ARIMA model (p, d, q), where in:

- p is the number of autoregressive terms;
- d is the number of differences; and
- q is the number of moving averages

RESULTS AND DISCUSSION

As this study is planning to prepare a best production model for spinach and find a best fitted model of Spinach production, based on different selected time series variables, such as Harvest area, average annual rainfall & temperature and market price whose data were retrieved from website https://www.fao.org/faostat/en/#data and climateknowledgeportal.worldbank.org. In a study, Shah et al. (2021) conducted an analysis to determine the various factors that influence the

yield of paddy crops. The researchers utilized time series data spanning from 2005 to 2014, specifically focusing on the Lodhran district in Pakistan. A multiple linear regression model was constructed in order to assess the impact of temperature and humidity on the yield of paddy rice. Several statistical tests were utilized to assess the adequacy of the model and to examine multicollinearity. These tests included R^2, adjusted R^2, Durbin Watson test, mean square error, P-value, and VIF. Similarly, this study has employed stepwise regression model to assess the best fitted multiple linear regression model (MLRM) to the dataset. The model's fitness and multicollinearity were assessed using R^2, adjusted R^2, residual standard error, P-value (Table no. 1) and VIF.
Considering these variables relative to the equivalent production level of spinach, the regression line for spinach production is as follows.

Spinach Production = -0.2322 + 0.0003 (HA) – 0.0015 (AMR) + 0.147 (AMT) + 5.984 (Price)

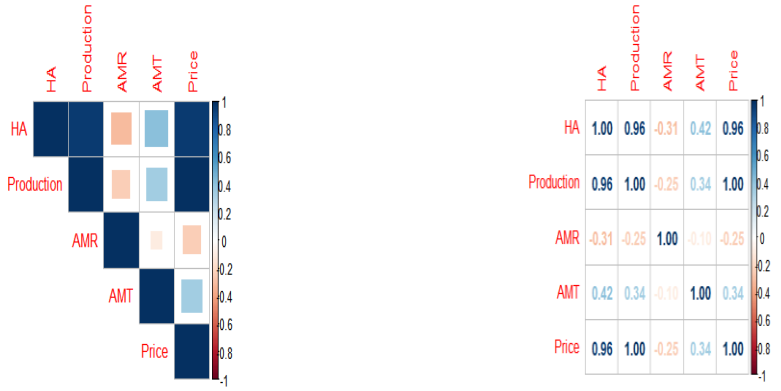
Table 1. Estimation and analysis of multiple linear regression models.

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	-0.2322	19.201	-0.012	0.990
HA	0.0003	0.0003	1.018	0.312
AMR	-0.0015	0.0009	1.686	0.097
AMT	0.1475	0.760	-0.194	0.846
Price	5.984	0.0003	19735.726	<0.001 ***
Residual standard error				1.7568
Multiple R-squared				Approximately equal to 1
Adjusted R-squared				Approximately equal to 1
F-statistic				$1.282e^{+09}$ (on 4 and 55 DF)
p-value				< 0.001

Correlation analyses

To check the multicollinearity problem as an assumption, firstly this study has tried to observe the correlation analysis between the independent variables (Fig. 1 (A)). Also, the values of simple correlation coefficient (r) between dependent and different independent variables are presented (Fig. 1 (B)). According to the data, there is a negative relationship between annual average rainfall (AMR) and spinach production, but favorable relationships between sales prices, harvest areas, and the annual average temperature. The correlation coefficient value is -0.25 between AMR and production which is weak negative correlation. The presence of correlation among the variables can be shown using the correlogram.

According to Figure 1, the correlation between harvest area and annual mean rainfall is -0.31, which postulates a weak negative relationship between the two variables. The correlation between the mean rainfall and temperature is -0.10, which indicates that there is either a weak negative association or nearly no relationship at all. Once again, rainfall has a weak negative linear connection of -0.25, but temperature has a weak positive linear link of 0.34 with price. But the harvest region likewise has a positive moderate linearity with an average temperature for the year of 0.42 and a significant connection with price of 0.96. This suggests that the two variables may be interrelated in some way. The correlation coefficient value is strong and very near to 1 between harvested area and price which is an indication of multicollinearity problem and any change in the value of the harvest area would result in a corresponding change in spinach price, leading to considerable fluctuations in the model results. The model's outcomes exhibit instability and significant variability when subjected to minor alterations in the data or model. So, to remove this multicollinearity problem we go through the stepwise regression for solution. In order to verify that the normalcy assumption is correct, descriptive statistics have been employed in Table 2. In this case, all of the components, with the exception of AMR, have a right-skew. It is evidence that none of the variables follow the normal distribution.



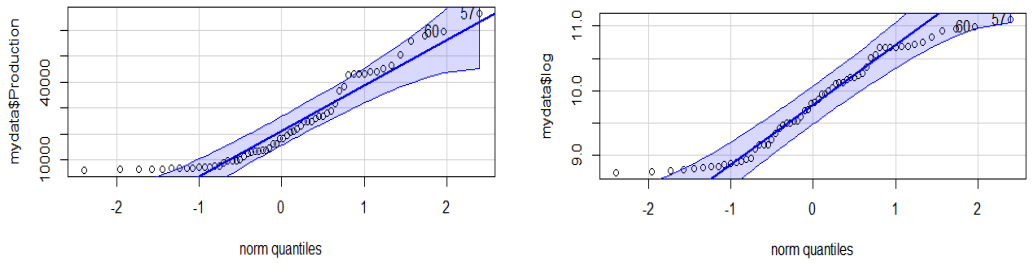
(A) (B)

Fig. 1. Correlogram of the different variables of Spinach production.

The following figure shows q-q plot for the Spinach production data and its log transforming data. First one graph is for Spinach production data and second one is from the log transformed data of the production. Original data shows better perform than log transformed data. First graph shows it slightly violet the normality of the data. Second graph shows a strong violation of normality assumption.

Table 2. Descriptive statistics.

Statistic	Harvest area	Production	AMR	AMT	Price
Mean	4399.22	23074.98	2245.0008	25.5795	3855.78
Median	3972.50	18257.50	2228.7550	25.5550	3051.00
Max	9580	66292	2844.49	26.58	11077
Min	1350	6200	1679.18	25.08	1036
Skewness	0.467	0.898	-0.103	0.860	0.898
Kurtosis	-1.147	-0.234	-0.587	0.525	-0.234



(A) (B)

Fig. 2. q-q plot for the spinach production and log transformed data.

Assumption checking

In spite of the fact that the overall model discovered significance in table 1, the vast majority of the determinants, with the exception of price and yearly rainfall, are not substantially started. So, this provides a suitable signal that the assumption of MLRM for this dataset has been violated. It would be an excellent moment to put the assumption of MLRM to the test. Residuals vs. fitted values, Q-Q plots, Scale location plots, and Cook's distance plots were used in order to assess the assumption of MLRM. These plots were also utilized in order to investigate the assumptions of linearity, normality, homogeneity, outliers, and high leverage points respectively.

Residual vs fitted values plot

The linearity assumption may be tested using graphs that compare residual values to values that have been fitted to the data. The red line on the plot ought to be horizontal when it's positioned at zero. The existence of a pattern can suggest that there is an issue with one of the components of the linear model. On our residual plot, there is no discernible pattern. So, it is reasonable to conclude, based on this plot, that there is a linear connection between the predictor factors and the result variables. This implies that a slight deviation from linearity may result in minimal practical implications. (Barker et al. 2015)

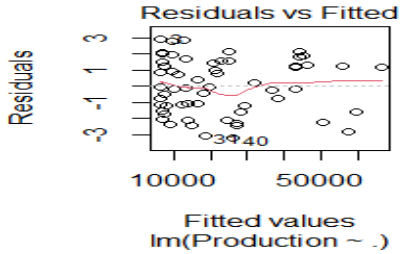


Fig. 3. Residual vs fitted values plot of spinach data.

Normal Q-Q plot

Via the use of the Residual plot, one may evaluate the normality assumption that was made. If everything goes according to plan, the normal probability distribution of residuals will look like a single direction. In this particular instance, the points cluster more or less along the boundary line, despite the fact that there are a few outliers. A perfectly straight line is not going to be possible. For non-normal data set there is an alternative method applied in Imon and Das, 2015.

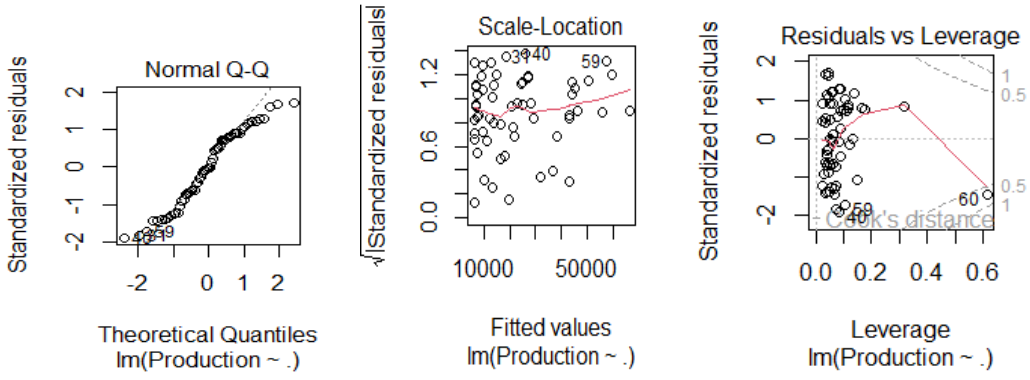


Fig. 4. Assumption checking different plots of spinach data.

Scale location plots

The Scale–Location plot is a useful tool for determining whether or not the assumption of homogeneity holds true. This graphic demonstrates whether or not the residuals are evenly distributed over the ranges of the predictors. It is desirable to be able to make out a horizontal line with points distributed evenly throughout its length. It can be observed that the variety of the residual points decreases with the value of the fitted outcome variable in this example, which suggests that the residual error variances remain constant. This can be seen by looking at the graph. At first glance, the residuals seem to be randomly distributed, despite the fact that there are outliers at the 40, 59, and 60 positions. The red smooth line is perpendicular to the x axis due to the fact

that the residuals are becoming more dispersed throughout a greater range. As a result, we do not face the challenge of heteroscedasticity in this situation.

Residuals vs leverage plots

Examining the residuals versus leverage plot will allow us to locate any outliers as well as high leverage points. The aforementioned plot shows the two most exceptional data points, specifically 59, and 60, which exhibit standardized residuals below -2. However, it is noteworthy that there are no outliers that surpass 3 standard deviations, which is considered favorable. Furthermore, it is evident that the data possess two high leverage points. All the data points in the sample have a leverage statistic that is below the threshold of 0.16, which is calculated as $2(p + 1)/n$, where p represents the number of predictors and n represents the sample size. Particularly the data point 60 poses high leverage value from the plot which is so far from the leverage threshold value.

Multicollinearity

After ensuring that the preceding assumptions are accurate, there is no completely traced plot over the data that can be used to fit the MLRM in order to make a forecast. In addition to that, the existence of significant multicollinearity has been shown. So, the subsequent phase will consist of determining whether or not multicollinearity exists using VIF. A greater number for the VIF suggests a higher level of multicollinearity. In this instance, the harvested area (HA) has a VIF of 14.654, which shows that there is the possibility of a strong correlation between HA and the other predictor variables in the model. In this particular instance, the coefficient estimates and p-values that are included in the result of the regression are probably not accurate. The value of the variable importance factor (VIF) for annual mean rainfall (AMR) is 1.158, which shows that there is a substantial connection between AMR and other predictor variables in the model; nevertheless, this is often not significant enough to call for attention. The value of the VIF for the annual mean temperature (AMT) is 1.277, which shows that there is no multicollinearity for the variable. Also the VIF value shows there is a strong correlation between price and other predictor variables in the model. As a rule of thumb, multicollinearity may be present if the Tolerance value is less than 0.1. For the variables HA and price this value shows an indication of multicollinearity problem.

Table 3. Variance inflation factor**.**

	HA	AMR	AMT	Price
VIF	14.654	1.158	1.277	13.168
Tolerance	0.068	0.863	0.783	0.075

The regression output, including the p-values and coefficient estimations, is very suspect. As can be seen in Table 1, the model's residual standard error is 1.748. This indicates that the

regression model can accurately forecast spinach yield with a standard deviation of 1.748. The results also show that the model has an R^2 of 1, which is an over-fitted model. Regression analysis shows that the significant F-statistic as the associated p-value is less than 0.001; therefore, the model is statistically significant. If the coefficient of determination (R^2) is one, then there is exact multicollinearity and in a multiple linear regression model, there is no information loss when one or more explanatory factors are removed from a set of variables exhibiting precise multicollinearity (Kim, 2019). In light of these facts, the strong multicollinearity between price and harvest area has resulted in a stepwise model being the most suitable fit for the data at hand. Consequently, the most efficient method for investigating the possibility of multicollinearity is a stepwise regression analysis.

Table 4. Eigenvalues and condition index.

Eigenvalue	Condition index	Intercept	HA	AMR	AMT	Price
4.59e+00	1.00	6.31e-06	0.0007	0.0005	6.16e-06	0.001
3.88e-01	3.43	4.04e-05	0.01	0.005	3.75e-05	0.02
1.40e-02	18.10	2.47e-04	0.51	0.148	3.20e-04	0.63
6.80e-03	25.98	3.31e-03	0.37	0.845	2.88e-03	0.29
6.89e-05	257.98	9.96e-01	0.09	0.0004	9.96e-01	0.04

Beside the VIF and Tolerance values, the eigen values and condition number is presented here to check the multicollinearity problem as this study finds R^2 value is exact 1. If the Condition Number is between 10 and 30, multicollinearity is likely present. Problems arise when values get beyond 30. Multicollinearity is strongly suggested by a high Condition Number and substantial amounts of variation as 0.50 or more. Multicollinearity is highly probable given the relatively high Condition Number (257.98) and the results of these three tests indicate that multicollinearity is present in our regression model.

Stepwise regression analysis

The optimal model for the stepwise procedure is the one with the lowest AIC. As can be shown in Table 5, Model 3 has the best AIC for predicting the variables. Model 3 was found to be the most accurate out of all the models that may be relevant based on the data we had. The lowest AIC value explains the greatest amount of variation using the price and rainfall data.

Table 5. Best subset regression of spinach production.

Model	Intercept	Price	HA	AMT	AMR	AIC
Model 1	X					1165.43
Model 2	X	X				69.74
Model 3	X	X			X	69.52

This meant the optimal model was the one that included the given predictors (harvest price and average yearly rainfall). The final model retains the inclusion of the relationship between price and rainfall, while excluding other variables. In the end, the "proper predictors" were singled out.

Table 6. An inventory of acceptable regressors.

SN	Predictor	Spinach (estimate)
1	Intercept	-3.051026
2	Price	5.984569
3	AMR	-0.001262

So, the final regression model for spinach is: $\hat{Y}$ = -3.051026 + 5.984569 Price -0.001262 AMR

While annual mean rainfall (AMR) remains constant, the equation predicts that a price increase will result in an additional 5.984569 times increase of spinach on average. It's an indication that the farmers' outlook on spinach production might improve if the market price of spinach's potential yield is bolstered. Once again, the output will decrease by 0.001262 times for every millimeter of more AMR (annual mean rainfall) on average if the other predictors remain constant. Also, the VIF values are 1.064 and 1.064 for price and annual mean rainfall in the final selected model which shows that there is no more multicollinearity problem in the final selected model now.

Prediction of time series data

In the decomposition of additive time series of Spinach production figures, it shows the same type pattern for the observed time series data and also for trend. To check the stationarity of the Spinach production data, Augmented Dickey-Fuller Test (ADF) applied here. In this test, the hypotheses are given below:

H_0: The time series is non-stationary.

H_A: The time series is stationary.

If the p-value from the test is less than some significance level (for example, =0.05), we can conclude that the time series is stationary. The ADF test statistic value for the Spinach production data is -0.747 with p-value 0.961. Hence, this test result shows this data is non-stationary. So, taking 1st differencing of the data, we can apply the ARIMA model. In this study, nine models are fitted as ARIMA (2, 1, 2) with drift, ARIMA (0, 1, 0) with drift, ARIMA (1, 1, 0) with drift, ARIMA (0, 1, 1) with drift, ARIMA (0, 1, 0), ARIMA (2, 1, 0) with drift, ARIMA (1, 1, 1) with drift, ARIMA (2, 1, 1) with drift, and ARIMA (1, 1, 0). Applying "forecast" package and "auto.arima()" function in R programming language, finally ARIMA (1, 1, 0) has been taken to forecast Spinach production in Bangladesh. The "forecast" package in R programming language offers a range of techniques and utilities for visualizing and examining predictions of univariate time series data. The auto.arima() function in the R programming language employs a modified version of the Hyndman-Khandakar method (Hyndman & Khandakar, 2008) to derive an ARIMA model. This algorithm integrates unit root tests, the minimization of the Akaike Information Criterion with a correction for small sample sizes (AICc), and maximum likelihood estimation (MLE) to determine the most suitable ARIMA model. The parameters of the auto.arima() function allow for a wide range of algorithmic changes. These include automated ARIMA modelling by "auto.arima()" function. ARIMA (1, 1, 0) is equivalent to having a partial autocorrelation function (PACF) of 1, a differencing of 1, and an autocorrelation function (ACF) of 0. Comparing among all nine models, it is found that ARIMA (1, 1, 0) with drift performs better as the AICc value is minimum (Table 7).

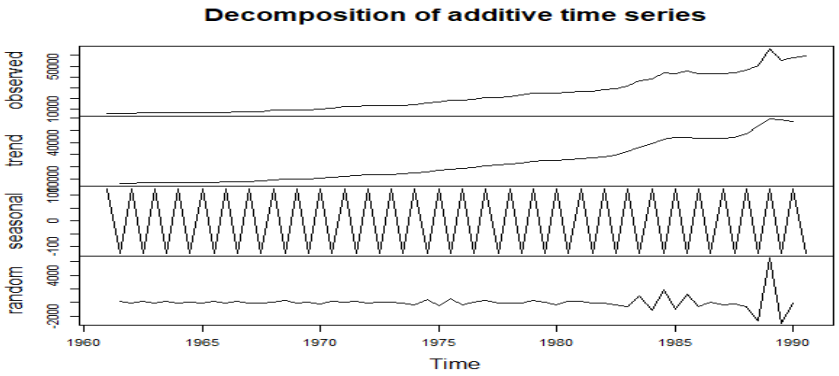


Fig. 5. Decomposition of additive time series of spinach production.

The model selection criteria like Akaike's Information Criteria (AIC), lowest Corrected Akaike's Information Criteria (AICc), Bayesian information criterion (BIC) values are presented below for the yearly spinach production data. The AIC and BIC values are 1103.55 and 1109.79 respectively for the ARIMA (1, 1, 0) with drift model. The estimated values are statistically significant (Table 8).

Table 7. Different models with AICc values.

Model	AICc
ARIMA (2, 1, 2) with drift	Inf.
ARIMA (0, 1, 0) with drift	1106.27
ARIMA (1, 1, 0) with drift	**1103.99**
ARIMA (0, 1, 1) with drift	1104.19
ARIMA (0, 1, 0)	1110.13
ARIMA (2, 1, 0) with drift	1106.29
ARIMA (1, 1, 1) with drift	1106.29
ARIMA (2, 1, 1) with drift	Inf.
ARIMA (1, 1, 0)	1110.93

Table 8. Parameter estimation of ARIMA (1, 1, 0) with drift model.

Parameter	Estimate	St. Error	z value	Pr (>|z|)
AR1	-0.269	0.124	-2.165*	0.0300
Drift	901.404	272.685	3.306**	0.0009

* and ** denotes significant at 5% and 1% level of significance

The accuracy function provides various metrics to evaluate the accuracy of the model fit, including mean error (ME), root mean squared error (RMSE), mean absolute error (MAE), mean percentage error (MPE), mean absolute percentage error (MAPE), mean absolute scaled error (MASE), and the first-order autocorrelation coefficient (ACF1). Here MASE value is less than 1 and also MAPE value looks good enough. Also ACF1 value is very far from 1 (in table 9). So the fitted model shows good accuracy and finally this model can forecast the Spinach production for upcoming years ahead. Figure 6 depicts the 15-year Spinach production forecasting and 95% confidence interval values based on the use of the autoregressive integrated moving average (ARIMA) (1, 1, 0) with drift model. In the graph, it shows the forecasted values of Spinach production by blue straight line. The deep blue and light blue regions show 80% and 95% confidence intervals of the forecasted Spinach production in Bangladesh. Almost 75,000 unit Spinach will be produced after fifteen years according to the ARIMA (1, 1, 0) model with drift. Also, the diagram shows an upward trend for the predicted production of Spinach in Bangladesh.

Table 9. Performance values of ARIMA (1, 1, 0) model.

Criteria	ARIMA (1, 1, 0)	Criteria	ARIMA (1, 1, 0)
Log likelihood	-548.78	RMSE	2626.28
$Sigma^2$	7260385	MAE	1328.84
AIC	1103.55	MPE	-3.35
AICc	1103.99	MAPE	6.42
BIC	1109.79	MASE	0.96
ME	-3.01	ACF1	-0.006

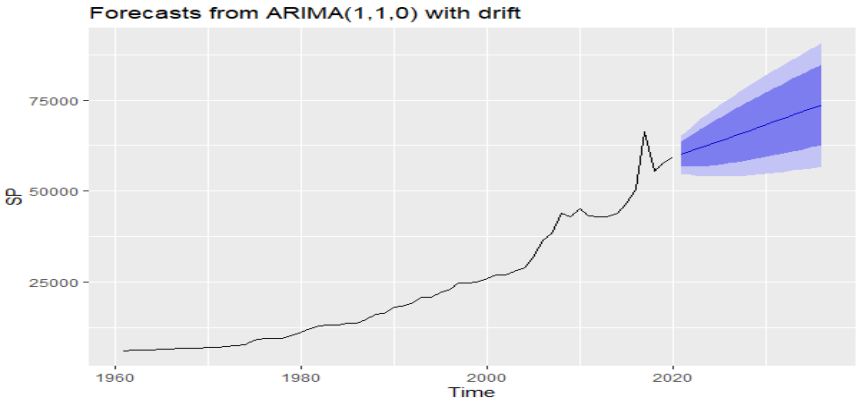


Fig. 6. Forecasting spinach production.

In their empirical study, Akhi et al. (2021) utilized time series data from 1961 to 2017 to forecast the production process of selected winter vegetables, including Bean, Cabbage, and Cauliflower, cultivated in Bangladesh. The researchers employed the Box Jenkins ARIMA methodology to analyze the production behavior and generate forecasts. The Box-Jenkins ARIMA models that provided the best fit for the data on Bean, Cabbage, and Cauliflower were estimated to be (0, 2, 1), (1, 2, 3), and (0, 2, 1) respectively. Das et al. (2022) applied six trend models in tea production time series data from 1976 to 2020. In this study, they found compound trend model is the most suitable model for the data set and forecast tea production. Our study utilizes the ARIMA (1, 1, 0) with drift model to forecast Spinach production in Bangladesh. The auto.arima() function in the R programming language is employed for this purpose. While certain analysis techniques utilized in this study can be found in other studies pertaining to diverse crop production forecasting,

it is challenging to locate any literature that has employed all of these techniques collectively within a single study for Spinach production. The study of production forecasting of spinach possesses distinctive characteristics.

CONCLUSION

As a nutritious and healthy food, the demand for Spinach is growing. So, it needs to analyze and forecast this production data to meet the future demand of Spinach in Bangladesh. Increasing future production will require careful selection of inputs. An appropriate model is selected in this study after checking performance and applying other diagnostic statistical methods. Also, a prediction is presented here for the future idea about Spinach production. Spinach production forecasts, in particular, have a significant impact on the future equilibrium between supply and demand. Government policy decisions concerning relative price structure, production, and consumption, as well as international relations, can benefit from these forecasts.

CONFLICT OF INTEREST

The authors declare that they have no conflicts of interest.

ACKNOWLEDGEMENTS

All the authors significantly contributed in the study. The authors express gratitude towards FAOSTAT and BBS for granting access to data from its Annual Report for the purpose of this study via its official website.

REFERENCES

Akhi, K., Sultana, N., & Sharmin, S. (2021). Production behavior and forecasting of some selected winter vegetables of Bangladesh, *Journal of Bangladesh Agricultural University, 19*(2), 251-260. https://doi.org/10.5455/JBAU.54123

Ali, M., Begum, A., Kakon, S., Karim, M. M., & Choudhury, D. (2022). Intercropping spinach and red amaranth with brinjal under different planting system. *Bangladesh Agronomy Journal, 25*(1), 91–96. https://doi.org/10.3329/baj.v25i1.62851

Babu, N. R., Divakar, J., Krishna, U., & Vigneshwaran, C. (2018). Study of antimicrobial, antioxidant, anti-inflammatory activities and phytochemical analysis of cooked and uncooked different spinach leaves. *Journal of Pharmacognosy and Phytochemistry, 7*(5), 1798–1803. https://www.phytojournal.com/archives/2018/vol7issue5/PartAE/7-4-667-208.pdf

Barker, L., & Shaw, K. M. (2015). Best (but oft-forgotten) practices: checking assumptions concerning regression residuals. *The American Journal of Clinical Nutrition, 102*(3), 533–539. https://doi.org/10.3945/ajcn.115.113498

Bergman, M., Varshavsk, L., Gottlieb, H. E., & Grossman, S. (2001). The antioxidant activity of aqueous spinach extract: chemical identification of active fractions. *Phytochemistry, 58*(1), 143-152. https://10.1016/s0031-9422(01)00137-6

Casson, R. J., & Farmer, L. D. M. (2014). Understanding and checking the assumptions of linear regression: a primer for medical researchers. *Clinical and Experimental Ophthalmology*, *42*(6), 590–596. https://doi.org/10.1111/ceo.12358

Chitwood, J., Shi, A., Evans, M. R., Rom, C. R., Gbur, E. E., Motes, D., Chen, P., & Hensley, D. L. (2016). Effect of temperature on seed germination in spinach (Spinacia oleracea). *Hortscience*, *51*(12), 1475–1478. https://doi.org/10.21273/hortsci11414-16

Das, K. R., & Imon A. H. M. R. (2016). A brief review of tests for normality. *American Journal of Theoretical and Applied Statistics*, *5*(1), 5. https://doi.org/10.11648/j.ajtas.20160501.12

Das, K. R., Sultana, N., Karmokar, P. K., & Hasan, M. N. (2022). A comparison of trend models for predicting tea production in Bangladesh. *Nepalese Journal of Statistics*, *6*, 51–62. https://doi.org/10.3126/njs.v6i01.50804

Frost, J. (2019). 5 Ways to Find Outliers in Your Data. *Statistics by Jim.* Retrieved from https://statisticsbyjim.com/basics/outliers/

Grossiord, C., Buckley, T. N., Cernusak, L. A., Novick, K. A., Poulter, B., Siegwolf, R., Sperry, J. S., & McDowell, N. G. (2020). Plant responses to rising vapor pressure deficit. *New Phytologist*, *226*(6), 1550–1566. https://doi.org/10.1111/nph.16485

Hasan, M. N., Rana, M. S., Malek, M., Das, K., & Sultana, N. (2016). Modeling Bangladesh's gross domestic product using regression approach. *Malaysian Journal of Mathematical Sciences*, *10*(2), 233–246. http://psasir.upm.edu.my/id/eprint/52343

Hyndman, R. J., & Khandakar, Y. (2008). Automatic time series forecasting: The forecast package for *R. Journal of Statistical Software*, *27*(3). https://doi.org/10.18637/jss.v027.i03

Imon, A. H. M. R., & Das, K. R. (2015). Analyzing length or size biased data: A study on the lengths of peas plants. *Malaysian Journal of Mathematical Sciences, 9*(1), 1-20.

Kim, J. H. (2019). Multicollinearity and misleading statistical results. *Korean Journal of Anesthesiology*, *72*(6), 558–569. https://doi.org/10.4097/kja.19087

Molyneux, P. (2004). The use of the stable free radical diphenyl picrylhydrazyl (DPPH) for estimating antioxidant activity. *Songklanakarin Journal of Science and Technology*, *26*(2), 211–219. https://rdo.psu.ac.th/sjstweb/journal/26-2/07-DPPH.pdf

Nath, C. R., Methela, N. J., Ruhi R. A., & Islam, M. S. (2020), Varietal performances of spinach (Spinacia oleracea L.) at coastal region in Bangladesh. *American-Eurasian Journal of Agronomy, 13*(2), 39-45. https://10.5829/idosi.aeja.2020.39.45

Naznin, M. T., Lefsrud, M., Gravel, V., & Azad M. O. K. (2019), Blue light added with red LEDs enhance growth characteristics, pigments content, and antioxidant capacity in lettuce, spinach, kale, basil, and sweet pepper in a controlled environment. *Plants, 8*(4), 93. https://10.3390/plants8040093

Osbourne, J. W., & Waters, E. (2002). Four assumptions of multiple regression that researchers should always test. *Practical Assessment, Research and Evaluation*, *8*(2), 2. https://doi.org/10.7275/r222-hv23

Pathania, A., Reddy, A. H., Rattan, P., & Kaur, S. (2022), Influence of planting time on yield and quality of spinach (Spinacia oleracea) varieties. *International Journal of Plant & Soil Science, 34*(23), 1174-1190. https://doi.org/10.9734/ijpss/2022/v34i232531

Seo Y, Ide K, Kitahata N, Kuchitsu K., & Dowaki K (2017). Environmental impact and nutritional improvement of elevated CO2 treatment: A case study of spinach production. *Sustainability*, *9*(10), 1854. https://doi.org/10.3390/su9101854

Shah, M. A. A., Mohsin, M., Chesneau, C., Sherwani, R. A. K., Fatima, A., & Bhatti, M. F. (2021). A statistical analysis on paddy crop production in southern Punjab, Pakistan. *Journal of Geography and Social Sciences*, *3*(2), 92-101.

Sharmin, S., Mitra, S. and Rashid, M. (2018). Production, yield and area growth of major winter vegetables of Bangladesh. *Journal of Bangladesh Agricultural University, 16*(3), 492–502. https://10.3329/jbau.v16i3.39447

Singh, R., Patel, C., Yadav, M. K., & Singh, K. K. (2014). Yield forecasting of rice and wheat crops for eastern Uttar Pradesh. *Journal of Agrometeorology*, 16(2), 199–202. https://doi.org/10.54386/jam.v16i2.1521

Tewani, R., Sharma, J. K., & Rao, S. V. (2016). Spinach (Palak) natural laxative. *International journal of applied research and technology*, *1*(2), 140-148.

World Bank (2007). *Project appraisal document on a proposed credit and a proposed loan to India for a National Agricultural Technology Project. Report No. 17082-IN.* The World Bank, Washington, DC.

World Bank (2008). *World development report 2008: Agriculture for development.* The World Bank, Washington, DC.

Reference to this paper should be made as follows:
Das, K. R., Jahan, M., Barman, L. R., & Burman, P. (2023). Modeling and forecasting of spinach production in Bangladesh. *Nep. J. Stat, 7,* 1-18.