National College of Ireland

Short term forecasting of Agro-products pricing using Multivariate time series analysis

MSc Research Project
MSc in Data Analytics

Raghav Krishna Kumar
Student ID: X18181848

School of Computing
National College of Ireland

Supervisor: Dr. Paul Stynes
Dr. Pramod Pathak

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Raghav krishna Kumar
Student ID:	18181848
Programme:	Msc in Data Analytics **Year:** 2019
Module:	Msc Research project
Supervisor:	Dr. Paul Stynes & Dr. Pramod Pathak
Submission Due Date:	17/08/2020
Project Title:	Short term forecasting of Agro-products pricing using multivariate time series analysis
Word Count:	8807 **Page Count:** 19

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Raghav krishna Kumar
Date:	17/08/2020

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Short term forecasting of Agro-products pricing using Multivariate time series analysis

Raghav Krishna Kumar
X18181848

Abstract:
Agro-products constitute to the daily needs of every person in this world. Frequent change in the prices of Agro-products create a huge pressure on the farmers and the consumers due to the money factor involved in cultivating the crops and buying daily food items. Determining and analysing the price fluctuation is a very complex study due to a large number of factors affecting it such as rainfall, temperature, holidays, air pollution and so on. Many researches have been done in the past to predict the commodity pricing but fails to explain the reliability factor and robustness of the model on the futuristic data. This research follows the designed framework to select the appropriate statistical/machine learning model based on data. This selects the model based on the data to forecast the future prices of Agro-products more accurately and increases the robustness of the model. Based on the framework Recurrent Neural Networks with Long-Short-Term-Memory is selected to predict the prices. Seasonal ARIMA and multiple linear regression models are also performed to compare the efficiency of the framework selected model. Based on the evaluations metrics and results of the three models, the RNN with LSTM model shows highly accurate results with a fit of 94% for predicting the prices of the Agro-products. This research project creates a model which helps the farmers and the consumers to identify the prices of the commodities for the future dates which helps them in avoiding any concurring losses.

Keywords: agricultural commodity price prediction, multivariate time series analysis, model selection framework, RNN with LSTM, Seasonal ARIMA, Multiple linear regression.

1. Introduction

Agricultural sector being one of the largest contributors to the world economy has at least 2 billion people working on it and rest of the population depending on it for their day to day life activities. The volatility in the prices of the Agro-products affect all the persons in this chain including the producers and the consumers. The fluctuation causes a huge disruption in the country's economy as the people are directly affected by the price increase. A recent example of price fluctuation could be seen in the prices of onion in India as there was a steep increase in the prices within a short duration which affected all the people's food habitat as onion constitutes to a major portion in daily food intake.

Many researches are being conducted in regard to determine the futuristic prices of the Agro-products from the early periods including the research by (Finkenstädt and Kuhbier, 1995) where the usage of non-parametric methods for non-linear time series data is performed with

the linear approximations technique. The research on the price prediction has been a long-time discussed topic due to continuous evolution in the prediction algorithms and advancement in the computer capacity. The evolution of neural networks in forecasting the prices is discussed by (Shahwan and Odening, 2007) where he uses the combination of traditional ARIMA model with the neural network. The hybrid model slightly increases the prediction rates compared to the ARIMA techniques. The latest research is done by (Dharavath and Khosla, 2019) uses time series analysis with Seasonal ARIMA which creates time lags that account for the seasonality factor in the time series. The results were able to predict the increase or decrease trend for the future dates.

The reliability and the robustness of the model plays an important role in the accepting the new futuristic data to the existing model. Most of the works in forecasting commodity pricing does not involve a methodology or provide the rationale behind choosing the models to predict. The trial and error techniques are used to select the model which thereby questions the reliability and the robustness of the model used for prediction purposes. This research project tries to overcome this limitation by using a framework to determine the time series model based on the data by using a series of tests to identify the seasonality, trend, causality, linearity and co-integration in the data.

The aim of this research is to investigate to what extent ML techniques can improve the accuracy of the daily price prediction of agricultural products using a model selection framework.

Based on the framework, RNN was selected and is compared with the latest work of (Dharavath and Khosla, 2019) Seasonal ARIMA and multiple linear regression which is used to evaluate the framework decision. Evaluating the selected model with one another shows that RNN model emerged as a superior model with low RMSE(root mean square error) and high fit percentage (R-Squared) compared to the other models specified. The inclusion of LSTM(Long-Short-Term-Memory) with the RNN architecture is being used improve the efficiency of the model in predicting the futuristic prices with high accuracy.

Based on the research question, following research objective were derived.

- Replicating the state-of-art model, Seasonal ARIMA used for predicting pricing of Agro-products
- Devising a framework for selecting the appropriate prediction model based on the data
- Implementing the RNN-LSTM and multiple linear regression model to predict the futuristic price of the commodities.
- Evaluating all the models with statistical metrics and comparing to determine the best fit model for prediction

The report is divided into different sections to make it clear to understand for the readers. The related works section discusses about the previous work and the state of art model for predicting the commodities prices. The research methodology explains the step by step process with accurate description of the research project. The design specification describes about the model and the framework used in this research project. The final stage of the model implementation is discussed in the implementation section followed by the evaluation section wherein different experiments conducted for the research project is compared. The results and discussion section

compares the different models performed to identify the most efficient model and the rationale behind it.

2. Literature review

2.1 Introduction:

The literature review section consists of the previous research papers related to the prediction of the agricultural commodity prices using various techniques and models. The section is differentiated by different sub-headings to give a clear overview of motivation of the project followed by the limitations on the previous models and enhancement done in this research paper.

2.2 Understanding the cause and effect of price volatility in Agro-products

The factors that influences the pricing of the agricultural prices and the effects caused by this volatility in the prices is discussed in this section with references to similar research papers. Researching on the volatility of the prices of Agro-products prove that a sudden increase or surge in the price creates a great impact on the middle-class families. As stated by (Birthal *et al.*, 2018) the volatility in prices are mainly caused due to demand factor, less production, inconsistency in import and export and poor supply chains within the country. The time series analysis of the price of onion reveals that the increasing demand in the market, failure in co-ordinating the supply chain management and the important factor is the perishable nature of the agricultural commodity which causes a huge decrease in the pricing when it nears its end stage.

The impacts and advantages of price discovery is studied in detail by (R L and Mishra, 2020)concluding that predicting the increase or decrease trend in the price of the Agro-products could help the producers to analyse the price formulation for the future investments. The implications of the price discovery is an important factor for the future markets which could help the producers and consumers to plan well ahead. (R L and Mishra, 2020) tries to strongly impose the point that the price fluctuation has been and will be a huge factor to determine the future of the agricultural sector in developing countries like India and this could be substantiated by the proof that even though the unemployment is growing higher year after year, the agricultural jobs are still in decline and the important reason is the uncertainty in the prices of the Agro-products. The effects of pricing volatility of agricultural products on a macro and micro economy is done by (Lanfranchi *et al.*, 2019). decrease in the import of agricultural products in third world countries could cause devastating effects as most of the country's GDP is directly linked to the contribution of the agricultural sector according to (Lanfranchi *et al.*, 2019) the impact in the micro-economy should be analysed using different socio-economic groups . Poor families and daily labourers are hit hard by the frequent fluctuations in the Agro-products. Some of the major causes and effects of the price fluctuation is discussed with reference to research papers which proves that there is major adverse impact of this fluctuation in the economy as well as the society. This research project tries to provide an alternate way to overcome the identified problem by predicting the prices of the Agro-products for the future days.

2.3 Selection of machine learning models for time series analysis

The process of selecting the appropriate models for prediction purpose has a great importance as failure to select the correct algorithm could drastically decrease the efficiency of the model. This section discusses the importance and the selection methodology of the predictive models.

Non-parametric methods which include the machine learning and deep learning models has a greater advantage in time series analysis as theory don't depend upon the prior distribution of the data and also the parameters involved with it as suggested by (Parmezan, Souza and Batista, 2019) due to this reason the non-parametric are easily adjustable based on different requirements and shows reliable results. the usage of different parametric estimation methods to select the parameters required for the models gives out the optimal value and best performance for the model. techniques like cross validation, box-Jenkins method are sufficiently used for estimating the optimal parameters for the machine learning models. the results produced by each model is compared with the models run without the optimal parameters shows greater efficiency in prediction of the dependent values. Among all other technique, Cross Validation technique used for estimating the parameters of machine learning model especially RNN for multivariate time series analysis shows greater accuracy in predicting the values compared to the other parametric model such as ARIMA and SVR.

This research project tries to provide the rationale behind the selection of the models based on the data by creating a framework which would determine the model based on a series of tests.

2.4 Finding an efficient machine learning model to predict commodity prices.

Many predictive algorithms and models have been developed since the rise of the modern computers with high computation power. This section reviews the previous research works on using different machine learning models and their advantages, drawbacks and how could provide inputs in the current research project.

(Jin *et al.*, 2019) uses LSTM deep learning technique on a time series analysis stating that neural networks work efficiently compared to the usual ARIMA statistical technique for prediction of cabbage prices in china. The use of STL for pre-processing the data has been used to eliminate the Seasonality and trend losses which improves the model's accuracy rate. RMSE and MAPE value are used to evaluate the accuracy of the prediction model on the monthly data of Agro-products. The impact of lag value has been identified and STL pre-processing technique is being used to eliminate. Even though the usage of RNN has given out a higher accuracy for the selected data, there is no comparison provided between ARIMA and RNN to substantiate the statement. The selection of algorithm has not been justified based on the data which could decrease the reliability factor of the model and decreases the robustness of the model.

(Zong and Zhu, 2012) in his work has used a unique technique called grey prediction to forecast the price of the agriculture products in china. RBF neural network is also implemented, and the results were compared with the grey prediction method using standard evaluation metrics like mean absolute percentage error. It was found that the grey method is not accurate in forecasting the prices when compared with the legacy neural network techniques. (Zhang *et al.*, 2020) has tried to forecast the price of the agricultural commodities using 29 different features extracted from various databases. There are models namely artificial neural networks, support vector

regression model and extreme learning machine were used, and the results were compared to arrive at the conclusion. In order to eradicate the redundancy, the minimum redundancy and maximum relevance approach is used. (Chen and Wang, 2007) in their work has used methods like spatial diffusion, microscopic models to discover the relationships between different parameters used to predict the price of the agricultural commodities. A real-world agriculture market is simulated, and the price of the agricultural products are used to help the farmers to determine the price. (Ye *et al.*, 2016) has tried to use complex recurrent neural network model with back propagation and ARIMA. ARIMA is a linear model and RNN with back propagation is non-linear. The output of this research shows that the non-linear models are better at forecasting the price of the agricultural commodities compared to the linear models like ARIMA and linear regression.

The usage of lasso regression method for predicting the futuristic price of the vegetables is proposed by (Yu *et al.*, 2019) and it substantiates the statement by comparing the results between the default OLS model with the lasso model. The regression co-efficient of lasso regression model fits the data with a higher accuracy thereby producing highly accurate results. The multicollinearity problem comes with the OLS regression as there are many factors included that influences the price fluctuation of the vegetables. the introduction of the lasso regression model tries to overcome the problem of multicollinearity by using the Mallows cp method. the explanation for the seasonality and lag was not done which could pose a series issue when the data is highly seasonal which could indicate the model following the seasonal pattern. The causality of the price of the vegetable with the other contributing factors has not been addressed. only multicollinearity of the factors among themselves does not provides information regarding the causality co-efficient between the dependent and the independent variables. (Chen *et al.*, 2019) suggest that using the wavelet analysis with LSTM technique could be used to predict the commodity pricing. The importance is given to the temporal information of the data and this forms the reason behind using the LSTM as this technique with RNN works good with long term data. The limitation of using RNN with LSTM with no external features to contribute to the independent variable, decreases the versatility of the research.

An attempt had been made to predict the tomato prices by (Ivanisevic *et al.*, 2015) using the ARIMA statistical model. The use of ARIMA model is justified by using different parametric estimation technique's which determines the appropriate values to be passed on. The model is a univariate analysis as it only has price which is a time dependent variable which provides the space to improvise the model with influential factor to make the model more robust. The usage of neural network models to predict the agricultural commodity pricing is done by (Anggraeni *et al.*, 2018) The MAPE evaluation metric is used to determine the efficiency of the model against the other models. The rationale behind the selection of hyper parameters used for artificial neural network model is presented clearly by evaluating it against a statistical metric. The optimal combination of values namely number of epoch's, the number of hidden layer and the output layer, learning rate of the algorithm are found using the iterative process with different combinations of the value which produces the least MAPE value. Even though the model's selection is not substantiated based on the data, the selection of the parameters is explained well. the model is built upon the optimal values of the search which produces a very low MAPE value. The selection of different evaluation metrics could explain the model's fit and efficiency in understanding the pattern of the data and predicting has been missed. The

seasonality factor as well as the causality factors has not been considered which makes the model lack the reliability factor.

A multivariate time series analysis using the support vector machine model with RBF kernel is attempted by (Shengwei *et al.*, 2018) and used the model to predict the futuristic prices of the Agro-products. The inclusion of external factors that influences the pricing of the commodities has been done to improve the efficiency of the model. the co-relation between the dependent variable which are the external factors in this case and the price of the agricultural commodity is found to eliminate the unnecessary factors that could wrongly influence the price rate. The meteorological factors such as the climatic conditions which could affect the model, the rate of consumption and other factors such as fertilizer rates and seed rates are considered in this model which makes the model more efficient in understanding the price fluctuation. the RMSE value of 0.1318 is achieved by the SVR model using RBF kernel which proves that the model is able to follow the trends introduced in the pattern. the amount of training data included foe the forecasting is very less as the model is trained using 2 years of data and tested on 2 months of data. the less amount of data used could also be the reason to achieve such a low RMSE value and this also poses a drawback as the model could fail when the futuristic data with lot of seasonality is fed to the model. Usage of smoothing techniques is suggested by (Rozahi Istambul and Abinowi, 2019) to predict the future prices of the chilli. Holt-winters smoothing technique is employed with seasonal multiplication method and is evaluate using the MAPE evaluation metric. Smoothing technique does not work efficiently with every set of data which reduces the reliability of the model for predicting the future data. The usage of long short time series network LSTNet is performed by (Ouyang, Wei and Wu, 2019) to predict the futuristic prices of the Agro-products. The model is compared with the traditional models used for prediction purposes such as ARIMA, CNN and VAR to evaluate the performance of the LSTNet. The selection of the horizon parameter is done by comparing the results of running with different value on different methods to finally land with a optimal value. the difference between the RNN and the LSTNet is found to be 6.52 in the root mean square error evaluation metric and RNN is found to work the efficient way compared to other traditional models. the inclusion of Long Short-Term Memory technique on top of the RNN could have increased the result to a much efficient way as both follow a similar technique to predict the price. the only drawback is that it lacks to explain when the data is in unobvious repetitive pattern and how could it be handled on a similar prediction scenario.

Seasonal ARIMA was used by (Paredes-Garcia *et al.*, 2019) to create a decision making system to forecast the commodity pricing. Seasonal ARIMA technique was used on a weekly based data to predict the future data. Seasonal ARIMA is used by (Dharavath and Khosla, 2019) to predict the futuristic prices of the Agro-products as a univariate time series analysis. The P,D,Q parameter estimation for the SARIMA model is acquired by the repetitive iteration method which gives out the optimal value with least AIC value. The model was able to recognise the seasonal trend of the prices of the different vegetables on a monthly basis. the rationale behind selecting the model based on the data was not explained which could decrease the model's robust nature as the model would more likely to be trained based on the seasonality and if the data does not contain the seasonal parameter, the model could be less efficient. Even though RMSE and MAPE values shows that the model is efficient in predicting the trend of the price in the future, the absolute and short-term determination of the prices could not be done with high accuracy.

Lot of models are discussed which are used for prediction purposes with their advantages and limitations from previous researches done. Based on the framework designed in this research project, the appropriate model is selected based on the nature of data.

2.5 Identification of gaps and conclusion

Based on the above literature review done on the existing papers, there is no established methodology in selecting the right prediction model as most of the research works are done with a prejudice assumption. Even though most of the models could produce the expected results with good results, the reliability factor of the model working on different data is very low. This gives rise to the research objective of this project where a framework is used to select the optimal model for prediction based on the data. The research tried to answer the question raised on how to improve the accuracy of the multivariate time series analysis for predicting the price of Agro-products.

3. Research methodology

This section of the report highlights the methodology followed from top to bottom of this research project in predicting the prices of the Agro-products. KDD approach which is a data analytics methodology has been followed in this research. The step by step process of research methodology is explained below.

3.1 Analysing Business Requirement:

The prices of the Agro-products cannot be made constant due to the price volatility nature. There are lot of factors which influence the price volatility such as the place of sale, production quantity, climatic factors etc., Understanding the requirements for predicting the commodities pricing plays an important role in the accurate forecast of the prices. The first requirement would be to gather the historical price of each of the agricultural commodity at a particular place. In this research project the prices of common vegetables and fruits such as onion, tomato, banana and cauliflower are collected for 10 years on a daily basis from 2006-2016 for the state of Delhi. The second part of requirement gathering is to acquire the values of the external factors that could influence the prices of these Agro-products such as rainfall, temperature, holidays, weekends, air pollution etc., Using these data the multivariate time series analysis is used to determine the prices of the selected commodity for the year 2016. It is then compared with the original values collected to measure the accuracy of the model.

3.2 Data Collection:

The historical prices of the Agro-products are collected from an open source government of India website[1] which lets the user extract the data based on different categories. The price details of onion, tomato, banana and cauliflower were extracted for 1 quintal with respect to Delhi and

[1] https://agmarknet.gov.in/

Azad market. The data for the external factors such as temperature, rainfall, air pollution was acquired from the climatic weather website of Indian government[2] for the same location.

3.3 Data Pre-processing and Data Transformation

The data acquired needs to be pre-processed and transformed to proceed with multivariate time series analysis. The null values from the dataset is handled by substituting the mean of the week thereby clearing the missing values in the dataset. The dataset is further segregated based on the market location inside the Delhi state and Azad market prices were considered for this multivariate time series analysis. The null value handling is also done for the external factors data with mean substitution. The data transformation activities such as changing the date according to required format, extracting different outputs using date-part function is done to make the data analysis ready.

3.4 Selection of Time series Model

Selection of the machine learning/statistical model to perform time series analysis plays a major role in the efficiency and accuracy of the prediction results. In this research project a framework has been used based on different test to determine the appropriate time series analysis model based on the data. The framework is used to analyse the stationarity, causality, linearity and co-integration properties of the data to select the time series Model. The flow chart is explained in the section 4(Design Architecture) and it is based on the results on each of the test performed to determine the perfect time series model.

3.5 Implementation of machine learning models

The data is now ready to be used for analysis and prediction. Selection of the machine learning model is determined with the help of the framework which is designed as part of this research. The framework follows a set of different tests to determine the data including the stationarity, linearity, causality and co-integration. Based on the results of the framework, the RNN methodology is selected for the particular data. The LSTM (Long-Short-Term-Memory) is a type of RNN which consists of a memory cell that could retain the cell information for a longer time period. Seasonal ARIMA which is the latest model implemented by (Dharavath and Khosla, 2019) is also implemented and evaluated to compare the results of the RNN-LSTM. Multiple Linear Regression is also implemented for the same data to evaluate the performance of the framework followed in this research project. The seasonal ARIMA explores the seasonality in the data and predicts based on previous seasonal trends. The model's efficiency is primarily determined by the P,D,Q parameters which is passed for the model to predict the futuristic prices. Grid search algorithm is used to determine the optimal value of the parameters which corresponds to a lowest AIC value.

The RNN-LSTM model which uses the principles of neural network to predict the values for the future date. The important difference between the SARIMA and RNN-LSTM is the consideration of external factors which influences the dependent variable. The external factors such as rainfall, temperature, holidays, weekends, air pollution are passed as the independent variable in this prediction. The parameters such as number of hidden nodes, the learning rate

[2] http://dsp.imdpune.gov.in/

and the number of epoch's are determined and are iterated to produce the least RMSE value. Multiple Linear Regression model works based on the linear relationship between the independent and the dependent variables. The feature selection techniques is used to find the most influential variable which affects the predictor variable and also used to eliminate the non-contributing features from the model.

3.6 Evaluating the models

The comparison between the different models used in predicting the agricultural commodity prices are evaluated using statistical evaluation metrics. The Root Mean Square Error reveals the error percentage between the actual values and the model predicted value. The R-Squared value determines the fit of the model on the data. The R-Squared value is always between 0 to 1 and value nearing 1 means good fit of the model. Visual representation of the actual and the predicted values gives the knowledge about the trends being followed by the predicted value compared to the actual values. Plots showing the actual vs predicted graphs provides an easy visualisation of the model's performance on predicting the future prices.

4. Design architecture

The selection of machine learning model is based on the below framework described in the figure 1. This framework consists of a series of test to determine the data and suggests the appropriate machine learning model which suits the data.

4.1 Stationarity check for the data:

Stationarity is an important factor to be considered when it comes to time series analysis. The time series without any seasonality and trend factor is considered to be a stationary time series. To check the stationarity of the data, dickey fuller test is used where the null hypothesis would be that the data is not stationary and alternate hypothesis is that the data does not have any trend or seasonality in it. This is the first step to access the data and if there is any trend or seasonality, the data needs to be differentiated to make it stationary.

4.2 Check for Granger-Casual relationship

The phenomenon of one time series causing or affecting the other time series is tested through granger causality relationship test. As this prediction is a multivariate time series analysis, there are more than one series of values that are time dependent. Hence checking based on the results of the casualty test should determine the second step in selecting the model. The granger causality test's null hypothesis is that the variable considered does not have any causal relationship with one another and the alternate hypothesis would be the vice versa. The model's VEC and VAR are selected if there exists a causal relationship and statistical/machine learning models are selected if there exists one.

4.3 Linearity check:

The linearity check for the data is required to select the appropriate model, as non-linear data are suitable for neural network method and linear data could be suited for ARIMAX and multiple linear regression based on the distribution of the data.

4.4 Co-Integration test:

The co-integration test is required to select the model if there is a granger causality exists. This is used to check if the non-stationary time series are integrated with one another. If the co-integration exists, then vector error correction model is selected and VAR (vector autoregressive model) is selected if there is no co-integration between the variables.

These tests are conducted on the data to determine the appropriate model based on the data. Further to this will be different parameter estimation methods based on the model selected.

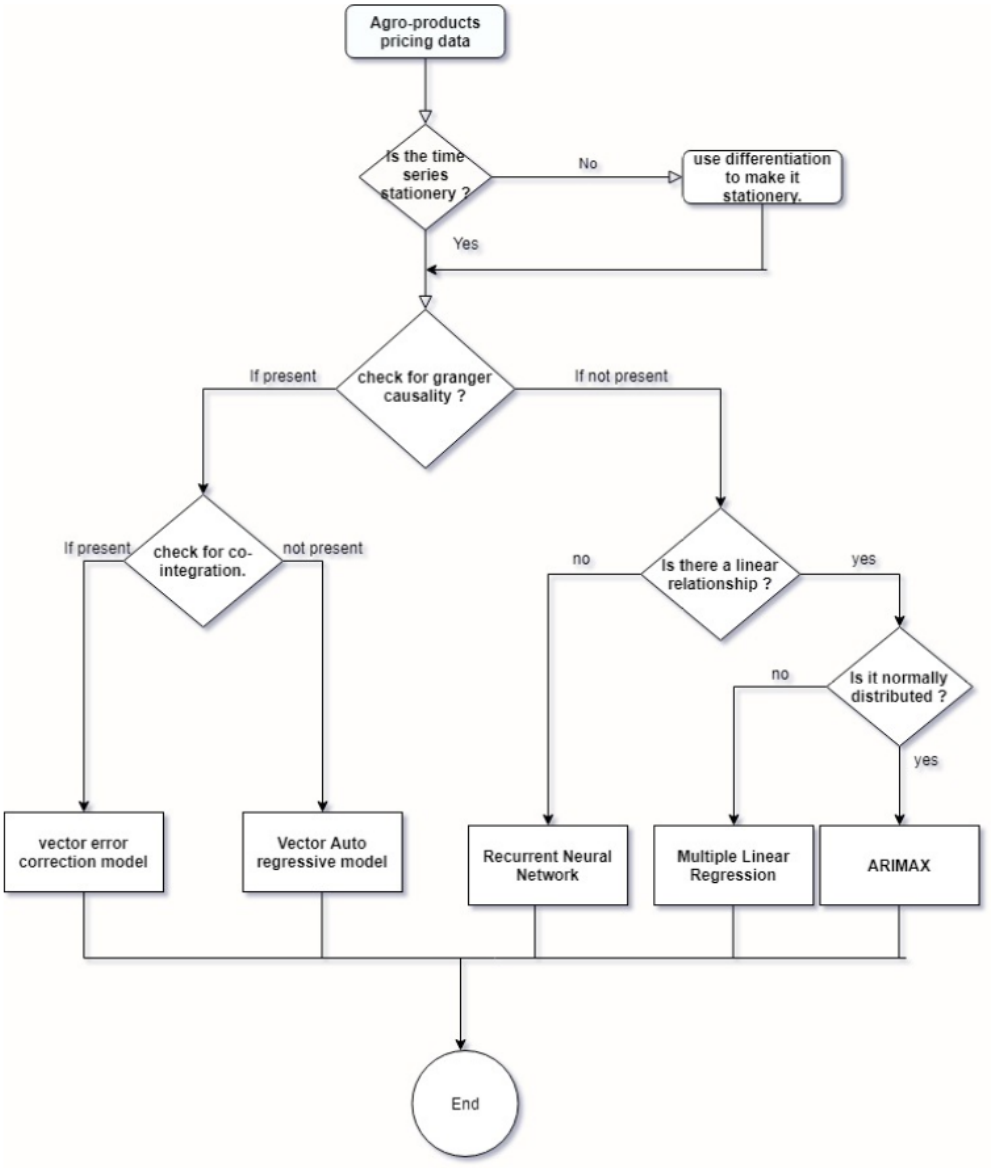

Figure 1: Framework to select the models.

6. Implementation

This section consists of step by step implementation of the research project based on the methodology section explained on how the data is processed, the model selection, fitting of the model and evaluating the results. This section follows a top to bottom approach in explaining the implementation of the research project.

6.1 Data Collection and Data Pre-processing:

The data is collected from an open sourced public repository organised by the government of India. The data was downloaded from the website in csv format which is later imported into python for model implementation. The null values of the data are handling using compensation by mean technique. The mean of the particular week for the particular market is calculated and the null values are substituted. The external factors which are collected as independent variables for the price prediction are extracted from the meteorological department of India which provides the daily climatic values based on the area. The other external factors such as holidays are extracted from the calendar function of the excel and is joined to the dataset using the VLOOKUP function in the excel. The datepart function from MySQL is used to determine the value of the quarter, weekends, day number of the year from the date value present in the main dataset. All the data are now collaborated using the excel and the final dataset is made analysis ready.

6.2 Selection of machine learning/statistical model :

The framework designed as explained in the section 4 form the base for selecting the model based on the series of the test and the results obtained from them. The first test which is stationary test was performed on the dependent variable(commodity price) and also on all the independent variables (temperature, rainfall, air pollution, weekends, holidays) and the results are analysed. The dicky fuller test is used to determine the stationarity of the data using the ADF package available in python. Figure 2 showcases the exploratory data plot to analyse the original and rolling mean of the data through which the stationarity is explained. The table 1 presents the information from the ADF test and based on the p-value the series is determined to be stationary.

Table 1: Result of ADF test

Results of Dickey-Fuller Test:	
p-value = 0.0060. The series is likely stationary.	
Test Statistic	-3.589308
p-value	0.005969
#Lags Used	5
Number of Observations Used	4126
Critical Value (1%)	-3.431936
Critical Value (5%)	-2.862241
Critical Value (10%)	-2.567143

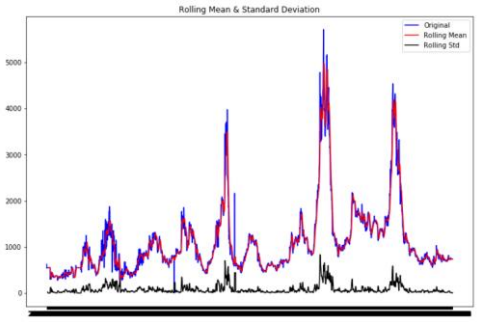

Figure 2: Framework to select the models.

According to the framework the next verification would be to test the granger causality test which is performed by using the stats model package importing the grangercausalitytest method. The maxlag parameter was set to 12 and the function tests for each lag to determine the optimal value. A matrix is formed between the variable of the dataset from which the row denoting the dependent variable is selected to check if the independent variables has a causality on the previous. The table 2 presents the values determined from the test and majority of the values are above the range of 0.05 which denotes and that majority of the variable does not have a causality on the dependent variable.

Table 2: Results of granger causality test.

	onion price	WSPD	vism	temp	rain	pressure	fog	hail	weekends	holidays
onion price	1	0.0101	0.0407	0.112	0.251	0.8716	0.1302	0.5456	0.0776	0.1052

Now we could eliminate the other flow of the framework thereby rejecting the usage of VAR and VEC models. The other flow consisting of the models related to the neural networks, ARIMA and multiple linear regression is selected and linear test is conducted to select the best working model among the selected three. The linearity test is conducted between the dependent and the independent variables and there is no linearity found. As per the framework neural networks will work best for this dataset and RNN-LSTM is used as it works efficiently with time series data .

6.3 Implementation of Recurrent Neural Networks:

The recurrent neural networks work best for the time series analysis as it understands the temporal sequence based on the input data thus eliminating the need to decode the lag observations as followed in traditional methods. The RNN requires the input data to be in a specific shaped array, so the data is first transformed into lagged observations T and T-1 based on the number of lag observations requested. The data is then split into test and train data based on the date. The data from 2006 to 2015 is split into the training set and the remaining one year of data is the test set. Both the dataset is then split into two which consists of dependent and independent variables. Then both the datasets are normalised as this makes the scale of all the dependent and the independent variable in the same range. The normalised data is transformed into a 3-dimensional array using the NumPy package in python. The 3-d input is then fed into

the RNN model to fit the model with the train data and predict it on the test data. The RNN model is built using the keras packages and with sequential as the base model. The number of epochs is set to 50 and the batch size is set to be 72. The loss value calculated is mean error and the activation function used is 'Adam' with the number of hidden layer set to 50. The trained model is then made to predict on the test set and the error rate RMSE and R-Squared is calculated to access the performance of the model. The value is demoralised again and is then plotted on the graph using matplotlib library with the original value. The error difference between the predicted and the actual values is calculated as RMSE and the fit of the model is determined through the R-squared.

6.4 Implementation of Seasonal Arima model :

The seasonal ARIMA is performed by for predicting the prices of Agro-products on a monthly basis. The data is collected is then converted into monthly values by considering the mean value of each month. The appropriate selection of P,D,Q parameters are required for the model to predict with high efficiency. The P represents the auto regressive part(AR), D represents the order of differencing(I) and the Q represents the order of the moving average and the same values are repeated for the seasonal component of the model with M stating the number of time steps. The selection of optimal values of these parameters are found through the box-jenkins method which uses the Auto-correlation plot to determine the Q value and partial auto-correlation plot to determine the P value initially. The ARMA model is fitted with initial value and the range for the iteration to be traversed is selected through the ACF and PACF plots drawn on the dependent variable. The initial values of P and Q is determined by the spikes found on the plots. the initial spike and the repetitive spikes determine the value of P which was 1 and the decay composition in the PACF was found to be 2 which represents the Q value. So, the initial value for the grid search was provided as (1,1,2). The range of the values are determined by the repetitive cycle obtained in the plots. So, the grid search range for d was given as (0,3) while the rest of the values are iterated in the range of (1,3). The optimal value based on the least AIC value was found to be (7,1,5) and using this value the ARIMAX model was run.

6.5 Implementation of Multiple Linear Regression :

Multiple Linear regression is carried to predict the agricultural prices by forming a linear relationship between the dependent and the independent variables, thereby predicting the futuristic prices based on the historic data. The variables in the dataset is checked for the linearity for performing linear regression. The data is then split into training and testing set and the dependent and independent variables are separated. The scikit library is used from which the linear regression package is called to fit in the dependent and the independent variables. The feature selection based on the co-efficient's is carried out to eliminate the non-influential variables from the model. The confidence level is set to 95% and the model is fitted with the training data. Then the predicted model is tested with the actual test data to determine the efficiency of the model. The RMSE value and the R-Squared values are obtained from the model to determine the error rate and the fit of the model.

7. Evaluation

7.1 Experiment 1: Replication of state-of-art model in predicting the commodities prices from historic prices using Seasonal ARIMA

Replicating the state-of-the-art model performed by (Dharavath and Khosla, 2019) forms the first experiment which helps to compare the other models used for prediction in this research project. The SARIMA technique is applied on the data with the parameters as P,D,Q values in SARIMA(0,1,2)x(2,1,1,3)and the RMSE, R-squared and the plot between the actual and the predicted has been used as evaluation metrics to determine the efficiency of the model. Table 1 presents the seasonal ARIMA method has predicted the prices of the commodity with a root mean squared error of 83.097 and a fit of 0.452. The root mean square error is the mean error difference between the predicted commodity prices and the actual prices from the dataset which means the lower the RMSE the higher the accuracy of the model. The R-Squared value is measured from 0 to 1 where 0 represents no fit and values nearing 1 will denote the perfect fit of the model in predicting the values.

Table 3: Evaluation metrics for SARIMA model

Model	RMSE	R-Squared
Seasonal ARIMA	83.097	0.452

The figure 3 compares the prices of the commodity between the actual and the predicted with respect to time series. The Y axis represents the price of the onion and the X axis represents the time duration. The time duration is denoted in number of days starting from JAN-2006 to DEC-2016 i.e., (01/01/2006 represents day 1 and 01/01/2007 represents 366) to give a clear image of the units. As the plot from figure 2 confirms that even though the error rate of the model is low, the model fails to identify the trends according to the real data.

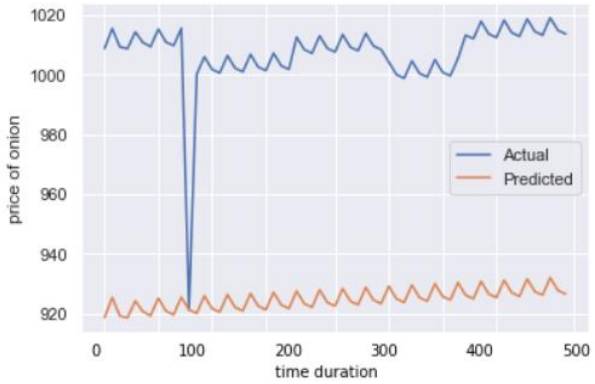

Figure 3: Predicted Vs Actual for Sarima Model.

As the Sarima model is replicated and does not give out good fit, the RNN-LSTM model is performed and compared with this model to compare and evaluate the best model amongst the both in the next experiment.

7.2 Experiment 2: Prediction of commodities prices using Recurrent Neural Networks with LSTM vs Seasonal Arima.

The Recurrent Neural Network-LSTM is used to predict the Agro-product pricing and the model is then compared with the Seasonal Arima(experiment1) to find the efficient model between the

both. The RNN-LSTM uses the same evaluation parameters used for evaluating the other experiments namely RMSE and R-Squared fit. The table 4 shows that the RMSE was found to be 30.85 which is the mean error difference between the predicted and the actual price. The R-squared is found to be 94.8 % which suggests that the models understands the data very well and is able to predict with high accuracy. The RNN-LSTM outperforms SARIMA model by a significant difference and the reason causing the difference is explained in section 8.

Table 4: Evaluation metrics for RNN-LSTM and SARIMA model

MODEL	RMSE	R-Squared
RNN-LSTM	30.85	0.948
Seasonal ARIMA	83.097	0.452

The plot between the actual and the predicted values is shown in the figure 4 which explains the trends followed by the values which are predicted by RNN against the original price of the Agro-products. The X axis units represents the day of the year as used for the SARIMA model.

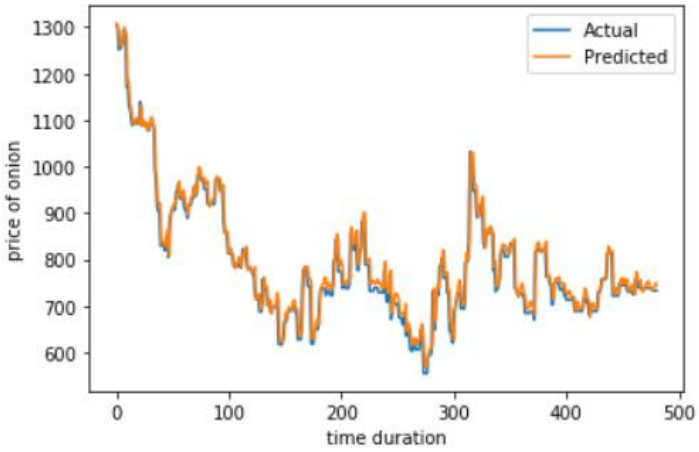

Figure 4: Predicted Vs Actual for RNN with LSTM Model.

The evaluation metrics and plots show that RNN-LSTM predicts with high accuracy compared to the SARIMA model. The need to test the efficiency of the model selection framework is carried out in experiment 3 between the model selected RNN-LSTM and Multiple linear regression which goes against the model.

7.3 Experiment 3: prediction of agricultural commodity prices using multiple linear regression vs RNN-LSTM model.

Multiple linear regression was selected to evaluate the framework used for selecting the model. Even though the data shows minimum amount of linearity between the variables, this model is implemented to shows the significance of the model selection based on the data. This gives a clear picture of how accurate the framework is in selecting the algorithm and producing efficient results.

Multiple linear regression predicts the Agro-products prices along with the independent variables namely temperature, rainfall, weekends and holidays. The model is then evaluated using the RMSE and R-Squared fit value. Table 5 explains the metrics obtained by the multiple linear regression model where the RMSE was found to be 273.93 which is the mean error difference between the predicted and the actual price. The R-squared is found to be -5.06 which suggests that the model is poorly fitted . When comparing the multiple linear regression results with RNN-LSTM , the RMSE value has been increased to a greater extent while the R-Squared fit is decreased.

Table 5: Evaluation metrics for Multiple linear regression vs RNN-LSTM

MODEL	RMSE	R-Squared
RNN-LSTM	30.85	0.948
Multiple Linear regression	273.93	-5.06

The figure 5 shows that the model is not able to identify the pattern and predict accurate results as the error between the actual and the predicted values are very high. This clearly states that this model did not outperform the RNN-LSTM which indicates the efficiency of the framework used for model selection.

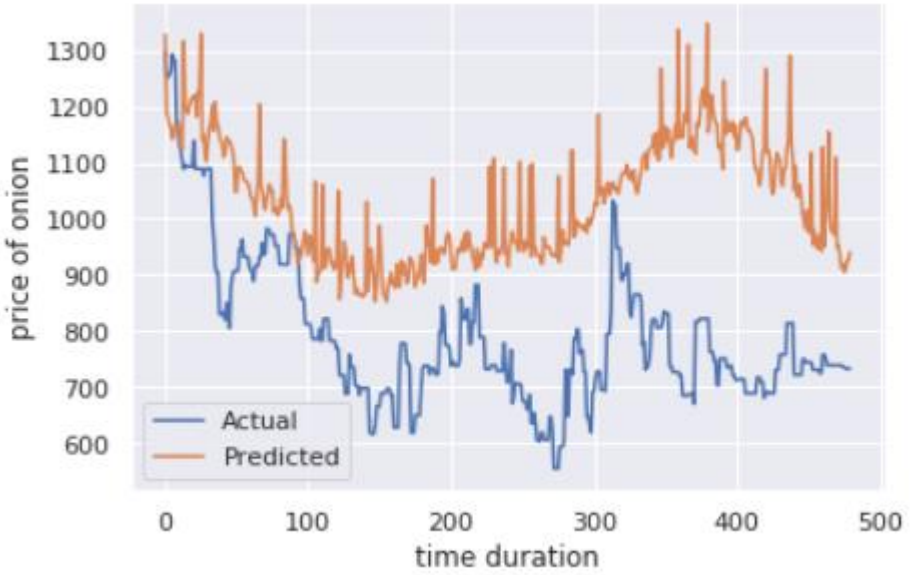

Figure 5: Predicted Vs Actual for Multiple Linear Regression

8. Results and discussion

The success of the model does not lie in predicting the expected output based on the historic data but to efficiently predict with the same level of accuracy for all the incoming and future data. A good model should be robust enough to handle the changes in the data yet produce the same results. When comparing the evaluation metrics such as RMSE and R-Squared, the RNN with LSTM model outperforms the other models by a huge margin. This huge margin of difference could be explained by the following reasons.

- selection methodology of Seasonal ARIMA model by (Dharavath and Khosla, 2019) which does not have a rationale behind choosing the model and could not

confirm if it is suited for the data. This could be explained by the poor R-Squared value from experiment 1 as shown in table 6. RNN-LSTM model overcomes this by performing exploratory analysis and following the data driven framework to determine the appropriate model.

- The Seasonal ARIMA model does a univariate time series analysis for a highly dependent price factor as there are lot of factors which influences the price determined for the commodities and failure to consider such factors does not make the model versatile enough to predict with high accuracy. The RNN with LSTM model uses multivariate time series analysis to includes factors that influences the predictor variable which makes it versatile enough to accommodate any future changes in the data.
- The multiple linear regression predicts the price of the vegetables based on the linear relationship between the dependent and the independent which makes it less prone to identify the varying pattern followed by the actual data as there was no linearity between the variables which explains a very high RMSE value as shown in figure 6. The model selection framework selects the model based on data and the low RMSE value and high R-Squared value of RNN-LSTM model proves that the framework selects the models accurately.

RNN model usually works well with short time data and fails with long-term data as it predicts based on the latest data considered. The LSTM work on the Recurrent Neural Network and overcomes the diminishing gradient problem usually faced by RNN and is capable of remembering long term dependencies to achieve state-of-the-art results in the time series prediction.

Table 6: Comparison of all the models

Models Implemented	**Root Mean Squared Error**	**R Squared**
Seasonal ARIMA	High	Not Accurate (0.452)
Multiple Linear Regression	High	Not Accurate (-5.06)
RNN with LSTM	Less	accurate(0.94)

9. Conclusion and future work

The price fluctuation of the Agro-products causes huge impact and adverse effects on the (farmers) and the consumers as it involves money loss. This research project tries to improve the accuracy of predicting the prices of the Agro-products for the future dates using multivariate time series analysis. A framework was followed in this research to select the precise time series prediction model based on the data. Based on the framework RNN-LSTM model was selected to be suited for the data considered. A comparison is made with the Seasonal ARIMA model to prove that RNN-LSTM works accurately with a R-Square of 94%. Multiple linear regression model was implemented to compare and check the accuracy of the model selection framework and the RNN-LSTM model predicts the price with a significant accuracy which states the success of the framework. This research project has contributed a valuable knowledge to the

farmers and public about the prices of the Agro-products well in advance which allows them to avert money losing situations and to avoid any further incurring losses.

The future work could be built on representing the visualisation and presentation of the results produced by the model more visually understanding by using various BI tools in a standalone or web application so that it could be easily available and accessible for the farmers and the public. Moreover, modern ETL tools like Alteryx could be used to automate this model by generating a pipeline which automatically pulls the daily data from the website and runs the model and pushes the output to the application.

References

Anggraeni, W. *et al.* (2018) 'Agricultural strategic commodity price forecasting using artificial neural network', in *2018 International Seminar on Research of Information Technology and Intelligent Systems, ISRITI 2018*. Institute of Electrical and Electronics Engineers Inc., pp. 347–352. doi: 10.1109/ISRITI.2018.8864442.

Birthal, P., Negi, A. and Joshi, P. K. (2019) 'Understanding causes of volatility in onion prices in India', *Journal of Agribusiness in Developing and Emerging Economies*. Emerald Group Publishing Ltd., 9(3), pp. 255–275. doi: 10.1108/JADEE-06-2018-0068.

Dharavath, R. and Khosla, E. (2019) 'Seasonal ARIMA to forecast fruits and vegetable agricultural prices', in *Proceedings - 2019 IEEE International Symposium on Smart Electronic Systems, iSES 2019*. Institute of Electrical and Electronics Engineers Inc., pp. 47–52. doi: 10.1109/iSES47678.2019.00023.

Ivanisevic, D. *et al.* (2015) 'Analysis and prediction of tomato price in serbia, *Ekonomika poljoprivrede*. Centre for Evaluation in Education and Science (CEON/CEES), 62(4), pp. 951–962. doi: 10.5937/ekopolj1504951i.

Jin, D. *et al.* (2019) 'Forecasting of vegetable prices using STL-LSTM method', in *2019 6th International Conference on Systems and Informatics, ICSAI 2019*. Institute of Electrical and Electronics Engineers Inc., pp. 866–871. doi: 10.1109/ICSAI48974.2019.9010181.

Lanfranchi, M. *et al.* (2019) *Economic and social impacts of price volatility in the markets of agricultural products*, *Bulgarian Journal of Agricultural Science*.

Parmezan, A. R. S., Souza, V. M. A. and Batista, G. E. A. P. A. (2019) 'Evaluation of statistical and machine learning models for time series prediction: Identifying the state-of-the-art and the best conditions for the use of each model', *Information Sciences*. Elsevier Inc., 484, pp. 302–337. doi: 10.1016/j.ins.2019.01.076.

R L, M. and Mishra, A. K. (2020) 'Price discovery and volatility spillover: an empirical evidence from spot and futures agricultural commodity markets in India', *Journal of Agribusiness in Developing and Emerging Economies*. Emerald Group Publishing Ltd. doi: 10.1108/JADEE-10-2019-0175.

Shengwei, W. *et al.* (2018) 'Agricultural price fluctuation model based on SVR', in *Proceedings*

of 2017 9th International Conference On Modelling, Identification and Control, ICMIC 2017. Institute of Electrical and Electronics Engineers Inc., pp. 545–550. doi: 10.1109/ICMIC.2017.8321704.

Yu, W. *et al.* (2019) 'Analysis of Vegetable Price Fluctuation Law and Causes based on Lasso Regression Model', in *Journal of Physics: Conference Series*. Institute of Physics Publishing, p. 012002. doi: 10.1088/1742-6596/1284/1/012002.

Chen, Y. J. and Wang, J. W. (2007) 'Agent-based simulation of agricultural prices volatility using cellular automata', *Second International Conference on Innovative Computing, Information and Control, ICICIC 2007*. doi: 10.1109/ICICIC.2007.121.

Ye, L. *et al.* (2016) 'Vegetables price forecasting in Hainan province based on linear and nonlinear combination model', *2016 13th International Conference on Service Systems and Service Management, ICSSSM 2016*. doi: 10.1109/ICSSSM.2016.7538566.

Zhang, D. *et al.* (2020) 'Forecasting Agricultural Commodity Prices Using Model Selection Framework with Time Series Features and Forecast Horizons', *IEEE Access*, 8, pp. 28197–28209. doi: 10.1109/ACCESS.2020.2971591.

Zong, J. and Zhu, Q. (2012) 'Apply grey prediction in the agriculture production price', *Proceedings - 2012 4th International Conference on Multimedia and Security, MINES 2012*. IEEE, pp. 396–399. doi: 10.1109/MINES.2012.78.

Chen, Q. et al. (2019) 'Price prediction of agricultural products based on wavelet analysis-lstm', in Proceedings - 2019 IEEE Intl Conf on Parallel and Distributed Processing with Applications, Big Data and Cloud Computing, Sustainable Computing and Communications, Social Computing and Networking, ISPA/BDCloud/SustainCom/SocialCom 2019. Institute of Electrical and Electronics Engineers Inc., pp. 984–990. doi: 10.1109/ISPA-BDCloud-SustainCom-SocialCom48970.2019.00142.

Ouyang, H., Wei, X. and Wu, Q. (2019) 'Agricultural commodity futures prices prediction via long- and short-term time series network', Journal of Applied Economics. Taylor and Francis Ltd., 22(1), pp. 468–483. doi: 10.1080/15140326.2019.1668664.

Paredes-Garcia, W. J. et al. (2019) 'Price Forecasting and Span Commercialization Opportunities for Mexican Agricultural Products', Agronomy. MDPI AG, 9(12), p. 826. doi: 10.3390/agronomy9120826.

Rozahi Istambul, M. and Abinowi, E. (2019) 'Prediction Red Chili Price Information in Bandung Use Smoothing Techniques', International Journal of Advanced Science and Technology, 28(6), pp. 146–152. Available at: http://sersc.org/journals/index.php/IJAST/article/view/385 (Accessed: 17 August 2020).

Finkenstädt, B. & Kuhbier, P. 1995, "Forecasting nonlinear economic time series: A simple test to accompany the nearest neighbor approach", Empirical Economics, vol. 20, no. 2, pp. 243-263.

Shahwan, T. and Odening, M. (2007) 'Forecasting agricultural commodity prices using hybrid neural networks', in Computational Intelligence in Economics and Finance: Volume II. Springer Berlin Heidelberg, pp. 63–74. doi: 10.1007/978-3-540-72821-4_3.