International Journal of Web-Based Learning and Teaching Technologies
Volume 15 • Issue 2 • April-June 2020

An Effective Multiple Linear Regression-Based Forecasting Model for Demand-Based Constructive Farming

Balaji Prabhu B.V., B.M.S College of Engineering, Bengaluru, VTU, Belgaum, India
M. Dakshayini, B.M.S College of Engineering, Bengaluru, VTU, Belgaum, India

ABSTRACT

Demand planning plays a very strategic role in improving the performance of every business, as the planning for a whole lot of other activities depends on the accuracy and validity of this exercise. The field of agriculture is not an exception; demand forecasting plays an important role in this area also, where a farmer can plan for the crop production according to the demand in future. Hence, a system which could forecasts the demand for day-to-day food harvests and assists the farmers in planning the crop production accordingly may lead to beneficial farming business. This paper would experiment by forecasting the demand using multiple linear regression (EMLR-DF) for different food commodities and implements the model to assists the farmers in demand based constructive farming. Implementation results have proved the effectiveness of the proposed system in educating the farmers in producing the yields mapping to the demand. Implementation and comparison results have proved the proposed EMLR-DF is more effective and accurate.

KEYWORDS

Agriculture, Business, Constructive Farming, Crop Production, Demand Planning, Forecasting, Multiple Linear Regression, Yield

1. INTRODUCTION

The farmers in developing countries like India are being faced by the age-old problems in the field of agriculture like there is no reliable and easy access to accurate weather forecasting, there is no easy access to government market portals, there is no common forum to consult the agricultural experts for discussion, there is no system to inform the farmers about new tools, technologies, and new governmental schemes, there is no synchronization between the demand and supply of food crops and list goes on. Demand-supply problem is one of the major problems being faced by the farmer community, where there is no synchronization in the production and demand for food crops. Due to this either farmer is failing to get good market prices when there is more supply than demand or consumer suffers high prices due to less production. To effectively tackle this problem, there is a need for some forum that forecasts the societal demand for different food crops and assists the farmers in growing the crops accordingly.

DOI: 10.4018/IJWLTT.2020040101

Demand forecasting plays an important role in the success of any business. As the demand forecasting helps in efficient planning for production, supply, maintenance, and other activity, the performance of a business depends on the accuracy of forecasting. Demand forecasting also plays a strategic role in the field of agriculture, where a farmer can plan for the crop production according to the need of society. So, this work attempts to develop a model to forecast the demand for the food crops and assists the farmers in selecting a crop based on this demand.

The demand for any product can be defined as, the aspiration for a product/commodity reinforced by the ability and willingness to buy it(Triplett, 1976). Demand is a function of Price of the Commodity P_c, Relative Price R_p, Income I, Taste and Preference T_p, Future Expectation F_e and other factors O ("Demand, Supply, and Market Equilibrium," 2011; Whelan & Forrester, 1996). The relation of Demand with these parameters can be expressed as:

- **Price of the Commodity (P_c):** Demand is contrariwise of the price i.e. the demand for a commodity will decrease with the increase in the price of a commodity and vice versa;
- **Relative Price of Commodity (R_p):** For certain products, the price are directly related to associated product price, those products are called relative products. The relative price of the product is directly proportional to the associated product;
- **The income (I):** The demand for a product will directly depend on the income of the people, an increase in income results in an increase in demand for the products. Consequently, the demand will decree with the decreasing income;
- **Taste and Preferences (T_p):** Taste and preferences of a consumers influence the demand for a commodity. With the change in taste and preferences of a consumer will affects the level of demand for various goods. Taste and Preferences may change due to the change in lifestyle, habits, etc.;
- **Future Expectation (F_e):** Expectation in the future price variations of goods will also affect their demand predominantly for durable goods. The expectation of an increase in price will decrease the demand and vice versa;
- **Other Factors (O):** Other factors such as population, weather and climate condition, taxation, advertisement etc. will also affect the demand of a commodity.

The Demand 'D' of the product could be represented as a function of the above parameters as $D= f (P_c, R_p, I, T_p, F_e, O)$.

But for some specific goods like food products, demand is not related to above-said factors and such goods are called necessary goods. such as food commodities like cereals, grains, vegetables etc. ("Demand Analysis and Forecasting," 2017). As the food commodities belong to the category of necessary goods, the demand for those commodities needs to be satisfied for the consumers by managing the supply accordingly. To regulate the demand and supply, demand for food commodities needs to be forecasted before and assist the farmers in planning their crop production accordingly. So that farmers can grow the food crops that would be in demand and caters the needs of the society. Demand forecasting for a product can be done using several methods (Ivano, 2017) such as Historical methods of forecasting which uses Opinion polling methods: consumer survey method, sales force opinion method and expert opinion method. Statistical methods could also be used to forecast the demand which uses trend projection method, barometric techniques, regression method, and simulations equation method. As the predictor variable 'Demand' is a numerical variable, Regression models are the best suitable for forecasting the demand values (Harlalka, 2018). Regression techniques are the most commonly used method in the process of demand forecasting (Hans, 2017).

Rest of the paper is organized as follows, the review of the existing works are discussed in section 2, section 3 gives a brief introduction to regression modes and section 4 explains the system architecture for the proposed demand forecasting model. The implementation and methodologies used in this work are described in section 5. Section 6 discusses the results and section 7 concludes the paper.

2. LITERATURE REVIEW

Crop selection is a critical problem being faced by farmers in the field of agriculture. Some of the researchers have worked on a selection of best crops based on the parameters like weather, soil, and yield rate of the crop etc. Rajesh Kumar et al. (R. Kumar, Singh, Kumar, & Singh, 2015) have proposed a crop selection method called (CSM) intended to increase the economic growth of the farmers based on yield rate, weather conditions, and soil conditions. The proposed method uses machine learning techniques to predict the yield rate, weather conditions, and soil conditions to check the suitability for crop selection. R Yadhav et al. (Rupika Yadav, Jhalak Rathod, 2015), have analyzed the problems with the traditional farming system and also proposed IoT and Big Data based precision agriculture system to improve the yields of the crops in farming. S Kothari et al. (Kothari, Channe, Kadam, & Professors, 2015) presented a cloud, IoT, Big Data based multidisciplinary model for smart agriculture. The proposed model will analyze the fertilizer requirement, best crop sequence based on the soil and weather conditions. The proposed system stores the soil data collected through sensors in the cloud for analysis. J Daniel et. al. (Jiménez, Dorado, Cock, Prager, & Delerce, 2016) have proposed a recommendation system for best crop selection based on the knowledge of the farmers and the data collected from the land using data analytics. D waga et al. (Duncan Waga, 2014) have considered environmental conditions like temperature, winds, rainfall etc. to analyze the suitability of different environmental conditions for different crop growing and to assist the farmers in best suitable crop selection based on environmental conditions. Elsheikh R et al. (Elsheikh et al., 2013) have developed a decision and planning tool evaluating the land suitability to choose an appropriate crop for the land based on the land specifications using GIS features. The Global Information System based systems were developed by (Nikkilä, Seilonen, & Koskinen, 2010; M.Narayana Reddy, 2017) to help the farmers in crop selection, fertilizer selection, water management and harvesting based on the analysis of GIS data. Prasad et. al. (Prasad, Peddoju, & Ghosh, 2016) have developed a mobile-based application called "AgroMobile", to help the farmers in identifying the crop diseases and the necessary actions needs to take for different disease through mobile app, also to help the farmers in marketing their harvest to the markets to gain better profits. Weather is a critical factor in crop farming, V Kumar et. al. (V. Kumar & Khan, 2017) have studied the need of weather forecasting in the best crop selection for improving the yield and also to manage the better crop cycles.

According to the survey made, most of the research works in the field of agriculture have focused on improving the yield of the crops, selecting the best crops based on weather condition and other land parameters. However, none of these works have concerned about producing the crops based on the needs of the society.

This paper investigates how forecasting the demand of various food commodities is performed using Multiple Linear Regression model and which could help the system to assist the farmers in cultivating the crops based on the demand so that the demand and supply of food crops could be mapped avoiding the loss for farmers leading to constructive farming. The various food commodities like Onion, Tomato, Rice and Wheat are considered to forecast the demand by the developed EMLR-DF (Effective Multiple Linear Regression based Demand Forecasting) model.

3. REGRESSION MODELS

Linear Regression is the simplest and most commonly used supervised machine learning method to build forecasting models. Regression models are used to forecast the value of a predictor variable given a set of explanatory variables by articulating the predictor variable as a function of explanatory variables. The relationship amongst the predictor variable and explanatory variables are represented in the form of a best-fit line with the minimum distance between the data points and the line or curve(Hans, 2017). The regression line is represented by a linear Equation (1):

$$Y = I + S*X + e \quad (1)$$

where:

'Y' is a predictor variable
'X' is the independent variable
'I' is the Intercept of the line
'S' is the slope of the line
'e' is the error term

Regression models can be categorized into Simple Linear Regression or Multiple Linear Regression based on the number of explanatory variables used to predict the value of the predictor variable.

3.1. Simple Linear Regression Model

A Linear Regression model where the predictor variable depends on a single explanatory variable is called a simple linear regression model. A simple linear regression equation can be represented as Equation (2):

$$Y = I + S*X + e \quad (2)$$

3.2. Multiple Linear Regression Model

A Regression model where the predictor variable depends on multiple explanatory variables is called as a multiple linear regression model (Higgins, 2005) (Simon, 2003). Multiple linear regression equations can be represented as Equation (3):

$$Y_i = I + S_1*X_{i,1} + S_2*X_{i,2} + \ldots.. + S_n*X_{i,n} + e_i \quad (3)$$

The slope and intercept values need to be calculated with respect to the explanatory variables to predict the Y values.

The slope of the equation can be calculated using the Equation (4):

$$\text{Slope(S)} = \left(N\sum X*Y - \left(\sum X\right)\left(\sum Y\right)\right) / \left(N\sum X^2 - \left(\sum X\right)^2\right) \quad (4)$$

The intercept of a line is calculated using the Equation (5):

$$\text{Intercept (A)} = \left(\sum Y - B\left(\sum X\right) / N\right) \quad (5)$$

where N is the total number of explanatory variables used in the model.

4. SYSTEM ARCHITECTURE

The demand of a product depends on various features("Factors of Supply and Demand," 2017) such as 'income of a person', 'sense of taste', 'price of a product', 'keen to buy a product' etc. However, the food produces belong to the group of essential products, the demand of food products mainly

depends on the factors such as 'demand of per capita', 'growth in income of a person', 'expenditure elasticity' and 'population' (Mittal, 2008). As the demand of a commodity D_c depends on multiple explanatory variables, the demand D_c for the period 'y' could be expressed as a function of explanatory variables as $D_c^y = f(D_{pc}, I_{pc}, E_e, P_y)$ where:

D_c^y = Demand forecasting for the commodity 'c' for the year 'y'
D_{pc} = Demand of per capita
I_{pc} = Income growth of per capita
E_e = Elasticity of expenditure
P_y = Projected population for the period 'y'

This demand equation can be expressed as linear combination of dependent (D_c^y) and independent (D_{pc}, I_{pc}, E_e, P_y) variables using regression equation (3.3) as a demand regression equation as shown in Equation (6). Where, the regression(Simon, 2003) is a supervised learning technique which tries to find a model from a dataset to generate a numerical prediction for future data samples:

$$D_c^y = \beta_0 + \beta_1 D_{pc} + \beta_2 I_{pc} + \beta_3 E_e + \beta_4 P_y + e \quad (6)$$

where β_0, β_1, β_2, β_3, and β_4 are the unknown parameters of the equation and 'e' is the error term.

Figure 1 shows the system architecture for the proposed EMLR-DF demand forecasting model, where the datasets containing the values of per capita demand for different commodities, growth in income, the expenditure elasticity for different commodities and the population growth for the previous ten years are stored in the database. The proposed regression model is applied to all the data sets stored and forecasted the demand for the specified period.

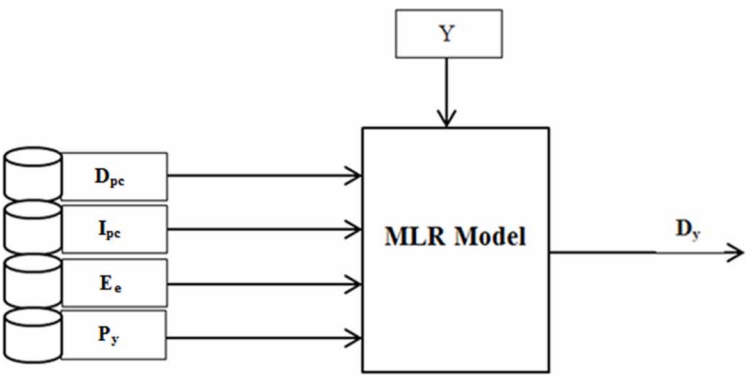
In general, this demand regression equation can be written as Equation (7):

$$Y_i = \beta_0 + \beta_1 * X_{i,1} + \beta_2 * X_{i,2} + \ldots.. + \beta_3 * X_{i,n} + e_i, \; i:=1,2,\ldots..,j \quad (7)$$

where j is the number of observations and n is the number of variables:

Y_i (i = 1,2, .., j) is the predictor variable- D_c^y
$X_{i,k}$ (k = 1,2, ..,n) are the explanatory variables used - D_{pc}, I_{pc}, E_e, P_y
β_0, β_1, β_2,……..,β_n are the unknown factors of the model
e_i is the term of random error, i=1,2,….n

Figure 1. Multiple linear regression based demand forecasting model



The values of any unknown parameters or coefficients could be predicted with the least square method. To simplify the calculation of model coefficients, the regression equation can be represented in matrix form as the matrix computation is more efficient than the ordinary least square computation method (João, Mzyece, & Kurien, 2009).

All the response values can represent in an n-dimensional vector which is called the response vector and represented as:

$$Y=\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

All explanatory variables can be represented as a '$n \times m + 1$' matrix considered as an independent matrix and represented as:

$$X=\begin{bmatrix} 1 & X_{11} & \cdots & X_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \cdots & X_{nm} \end{bmatrix}$$

The intercepts and slopes can be put into a p+1 dimensional vector called the slope vector, and represented as:

$$S=\begin{pmatrix} I \\ S_1 \\ \vdots \\ S_m \end{pmatrix}$$

Finally, all the errors terms can be written into an n-dimensional vector called the error vector and represented as:

$$e=\begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

The equation of the multiple linear models:

$$Y_i = A + B_1 * X_{i,1} + B_2 * X_{i,2} + \ldots.. + B_m * X_{i,m} + e_i$$

can be written as:

$$Y = A + X * B + e$$

where X*B is the matrix-vector product.

So, the Multiple Linear Regression Equation (7) can be written as:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{pmatrix} 1 & X_{11} & \cdots & X_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \cdots & X_{nm} \end{pmatrix} \begin{bmatrix} 2 \\ 0 \\ 2 \\ 1 \\ \vdots \\ 2 \\ m \end{bmatrix}$$

i.e y = X β, where $\beta = \left(X^T X\right)^{-1} X^T y$, where X^T : Transpose matrix X.

Further, the computation of coefficients can be simplified by decomposing the data matrix using any matrix decomposition method (João et al., 2009). QR decomposition(Agarwal & Mehra, 2014) is one of the most commonly used methods in matrix decomposition to solve the ordinary least square problem. QR matrix decomposition is a decomposition of a matrix X into an orthogonal matrix Q and an upper triangular matrix R as:

X = Q R

$$\begin{pmatrix} X_{11} & \cdots & X_{1n} \\ \vdots & \ddots & \vdots \\ X_{n1} & \cdots & X_{nm} \end{pmatrix} = \begin{pmatrix} Q_{11} & \cdots & Q_{1n} \\ \vdots & \ddots & \vdots \\ Q_{n1} & \cdots & Q_{nm} \end{pmatrix} \begin{pmatrix} R_{11} & \cdots & R_{1n} \\ \vdots & 0 & \vdots \\ 0 & 0 & 0 \end{pmatrix}$$

In $\beta = \left(X^T X\right)^{-1} X^T y$ replacing X with QR results in $\beta = R^{-1} Q^T Y$.

This computation method will reduce the number of computations and also eases the coefficient calculation. Nonetheless, in this method, the increased number of observations of independent variables makes the computation more complex for the multiple linear regressions. To overcome this problem, in this work, a MapReduce based ELR-DF model is considered. Where the huge computations are distributed among multiple clusters of nodes and calculated in parallel and independently. Hence, this approach reduces the complexity of coefficient calculation significantly for multiple linear regressions. The execution time required by the parallel multiple linear regression method is very less compared to the least square method of computation and matrix decomposition method. The results show the performance of parallel multiple regression models vs. least square and matrix decomposition methods.

5. IMPLEMENTATION AND METHODOLOGY

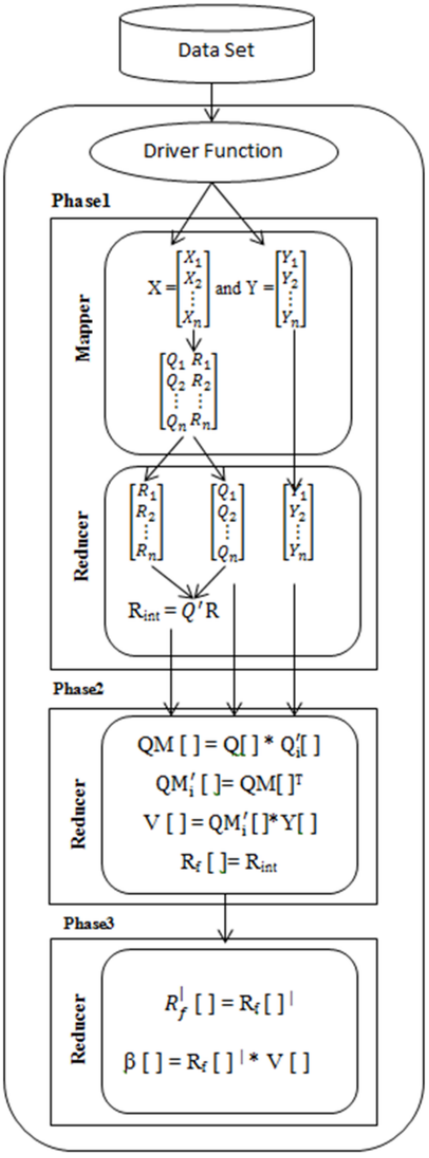
The proposed model is implemented in R-Hadoop environment using the Map-Reduce paradigm. The datasets collected are stored in a distributed mode using HDFS to make the data processing parallel. The proposed demand forecasting model is implemented in Map-Reduce to provide the scalability. Map phases split the data into different partitions and produce the intermediate results in to reduce phase. The proposed algorithm is implemented in reduce phase to compute the model coefficients. The demand is forecasted using the computed coefficients.

The data sets required for the demand forecasting such as values of per capita demand for different commodities, growth in income, the expenditure elasticity for different commodities and the population growth for the previous ten years are gathered from the 'Agmarknet' ["Agmarknet," 2018], NSS[Government of Karnataka, 2012, India National Sample Survey Office, 2010], and other authorized government websites.

Figure 2 gives the flowchart representation for coefficients calculation using parallel multiple regression model. The regression model has been implemented in three phases using the MapReduce paradigm. The driver function has divided the input data into X(independent variables) and Y(dependent variable) matrix and generate as an input to mapper in phase1.

The Mapper in phase 1 decompose the X matrix into Q and R matrix using QR decomposition method and produce Q and R matrix as an input to the reducer. The reducer in phase1 uses the R and Q matrix and computes the multiplication of Q inverse and R matrix as intermediate results to phase2. The reducer in phase2 considers the Q matrix and computes the inverse of Q matrix and multiplies it with the Q matrix and stores the result in QM matrix by transposing it. The transpose of the QM matrix is multiplied with the Y matrix and the resultant matrix V is produced as an input

Figure 2. Flowchart representation of the parallel multiple linear regression model



to phase3. The reducer function implemented in phase3 will multiply the matrix V with the inverse of matrix R and computes the model coefficients.

5.1. Parallel Multiple Linear Regression Model

Multiple linear regression algorithm has been implemented in Map-Reduce in three different phases.

5.1.1. Phase 1

Driver function: The driver function takes the parameter "Block size" which is used to divide the huge datasets related to the independent variables X_i and dependent variable Y_i into the blocks of small partitions when uploaded to HDFS. Each partition will be sent to a Mapper:

$$X(m,n) = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} \text{ and } Y(m,n) = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$$

Driver function Algorithm

```
Input: Dataset D
Output: X_i, Y_i; where i=1….n partitions.
[N_b=Number of blocks, BS=Block size,
    Start
       X= D [1:4]    //range of independent observations
       Y= D [5]      //range of dependent variable
       N_b = Size(X) / BS
       X_i = Block_Divide(X, BS)
       Y_i = Block_Divide(Y, BS)
    End
```

Mapper function: Every mapper job receives the X_i and Y_i matrix from driver phase. The matrix X_i contains the values of per capita demand for different commodities, growth in income, the expenditure elasticity and the population growth for the previous ten years. The matrix Y_i contains the values of the demand for the food commodities in the base years. The input vector Y_i is sent directly to reducer with the key <Key_i>. QR decomposition function has been implemented in each map job using Map_Fact() function to decompose the received matrix X_i into Q and R matrix. The map function produces a matrices Q and R with the key <Key_i> and <Key_R> respectively and sent them to the reducer. The input matrix **X** is decomposed into QR matrix as:

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} = \begin{bmatrix} Q_1 & R_1 \\ Q_2 & R_2 \\ \vdots & \vdots \\ Q_n & R_n \end{bmatrix}$$

Mapper function Algorithm [QR Decomposition]

```
Input: X_i, Y_i; where i=1,…n Blocks
Output: Q_i, R_i; Where i=1….n Blocks
   Start
         (Q_i,R_i)= Mapp_Fact(X_i)
         Produce(Key_i, Q_i)
         Produce(Key_R, R_i)
         Produce(Key_i,Y_i)
   End
Map_Fact(X_i)
Input: X_i
Output: Q_i,R_i
   Start
         (Q_i,R_i)=QR_Fact(X_i)
   End
```

Reducer function: A single reduce task has been implemented in phase 1 to process the results received from the map phase. The reducer function will receive inputs as the set of R_i, Q_i matrices and Y_i vector from the map function. The Q_i matrices and Y_i vectors are sent to the reducer in the second phase. The matrix R_i is used by this reducer to compute the intermediate matrix R_{int}. The intermediate matrix R_{int} is then decomposed using QR factorization as:

$$R_{int} = \begin{bmatrix} R_1 \\ R_2 \\ \vdots \\ R_n \end{bmatrix}$$

R_{int} is then decomposed using QR factorization as:

$$R_{int} = Q'R$$

The matrix Q' will be decomposed in to the same small partitions using reduce_1() function and sent to the reducer in phase 2 with the key <Key_i>. The matrix R is consider to construct the final matrix R_f with the key <Kay_R>.

Reducer function Algorithm [QR Factorization]

```
Reducer()
Start
     Reduce_1 (Key_i, List[Y_i, Q_1,Q_2,..Q_i])
     Input: Q_i, Y_i, where i=1…n Blocks
     Start
          Produce (Key_i, Y_i)
          Produce (Key_i, Q_i)
     End
     Reduce_1(Key_R, [R_1,….R_i])
     Input: R_int = Matrix [R_1,…R_i]
     Output: = Q_i' and R_f ; Where i=1…n Blocks
```

```
      Start
       (Q′, R_f)=QR_Factorize(R_int)
        Q_i^' = Block_Divide(Q′,BS)
      End
End
```

5.1.2. Phase 2

A single reduce task has been implemented in phase 2. The input to this reduce function is the set of $Q_i^{'}$, Q_i and Y_i matrices which have the same key <Key$_i$> from the phase1. The reducer in phase2 multiply the corresponding ($Q_i^{'}$, Q_i) factors and store the resultant matrix into QM matrix. The resultant matrix QM is transposed and multiplied with the Y_i vector to produce the intermediate vector V_i.

Reducer function Algorithm [Intermediate values]

```
Start
       Reducer_V (Key_i, List[Q_i, Q_i^', Y_i])
              Input: List [Q_i, Q_i^', Y_i)
              Output: = V_i[ ]
              Start
                     QM_i = multiply(Q_i, Q_i^')
                     QM_i^' = Transpose(QM_i)
                      V_i = multiply(QM_i^', Y_i)
              End
       Reduce_Key (Key_R, R_f)
              Input = R_f
              Output= R_f
              Start
                     Produce(Key_f, R_f)
              End
End
```

5.1.3. Phase 3

A single reduce task has been implemented in phase 3 to produce the final result. The input to reduce function is the set of vectors V_i and R_f from phase 2. The addition of all V_i vectors gives the final vector V, and the final vector V is multiplied with the final R value R_f to produce the regression coefficients β values.

Reducer function Algorithm [Coefficients Calculation]

```
Start
        Reducer_B(Key_f, List[R_f, V1,…Vi])
        Input: List[R_f, V1,..Vi]
        Start
                R_f^{-1} = Inverse(R_f)
                B_i[ ] = multiply(R_f^{-1}, ∑V_i)
        End
        Output: B_i[ ]
End
```

The computed coefficients (β) values are applied to the regression equation to predict the demand for the selected crop for the considered Harvesting period.

6. RESULTS AND DISCUSSION

In order to build an accurate demand-based forecasting model, it is feasible to consider a set of independent (response) and dependent (predictor) variables for MLR method which are linearly related and strongly correlated among themselves. So, it is necessary to check whether there exist any linear relationship and significant correlation between the dependent and independent variables considered. The relation between the response and predictor variables can be analyzed through computation of the simple summary function with means standard deviation, correlation etc. The relationship can be visualized using the scatter plots (Figures 2 and 3), which shows the existence of a linear association between the response variables and predictor variables, outliers, data-entry errors and skewed or unusual distributions. Before implementing the proposed EMLR-DF model, the efficiency could be evaluated with the sample data sets for the crops Rice and Onion considering demand as dependent variable and population, income, per capita demand and expenditure as independent variables using correlation coefficients, R-squared values and scatter plots as the metrics.

Figure 3 is the scatter plot for analyzing the relationship between the dependent and independent variables used in the demand forecasting model for the crop Rice. From the pattern in which data points are scattered in the figure, it can be determined that there exists a linear relationship between the variables of the model. So, the preferred variables could be used in implementing the model.

Table 1 also shows the correlation between each and every variable of the model in all possible combinations which are very strong as on an average, it is 89%. As all the independent variables population, income, expenditure and per capita demand exhibits a strong correlation with the dependent

Figure 3. Scatter plot – Linear relationship between dependent and independent variables while forecasting the demand for the crop rice

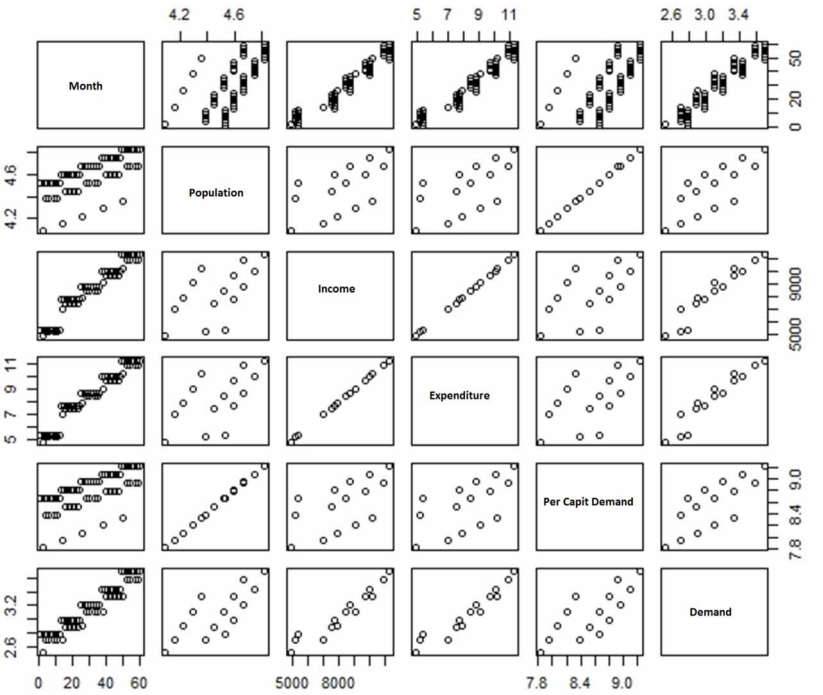


Table 1. Correlation between dependent and independent variables

	Population	Income	Expenditure	Per Capita Demand	Demand
Population	1.0000000	0.7040245	0.7040245	0.999792	0.812162
Income	0.7040245	1.0000000	1.0000000	0.692819	0.969624
Expenditure	0.7040245	1.0000000	1.0000000	0.692819	0.867482
Per capita demand	0.9997917	0.6928186	0.6928186	1.000000	0.801281
Demand	0.8121618	0.9696239	0.9696239	0.801281	1.000000

variable demand, these independent variables could be certainly considered as more suitable and significant towards the development of the model. The level of correlation or the significance of these coefficients is shown in Table 2 and observed that all the variables are highly significant(***) with the developed model.

The proposed model gives the most significant values with the more significant correlation coefficients and the R-squared value of 0.98 that indicates 98% of accuracy by the developed model in predicting the demand values.

Table 3 compares the actual demand value and the model predicted demand values for the crop rice. From the table, it can be observed that the predicted values by the model are on par with the actual values. So, the developed model can effectively forecast the demand values for food crops. So the developed model could be used in forecasting the demand values for agricultural commodities very effectively.

Figure 4 is the scatter plot for analyzing the relationship between the dependent and independent variables used in the demand forecasting model for the crop Onion. From the figure, it can be determined that there exists a linear relationship between the variables of the model. So, the preferred variables could be used in implementing the model. Table 4 shows the correlation between the variables of the model, from the table it can be observed that there exist a strong correlation between the variables of the model. As the independent variables exhibit a strong correlation with the dependent variable, used independent variables are more significant in model development. Table 5 gives the significance of model coefficients and shows the high significance (***) values for all the variables chosen.

The Model coefficient value shows the high significance with the 'p' value and R-squared values of 0.99 that indicates 99% of accuracy by the developed model in predicting the demand values. So, the model could be used effectively in forecasting the demand.

From the analysis made with the results of scatter plot and correlation function, it can be concluded that the proposed EMLR-DF model is the most appropriate in forecasting the demand for the food crops

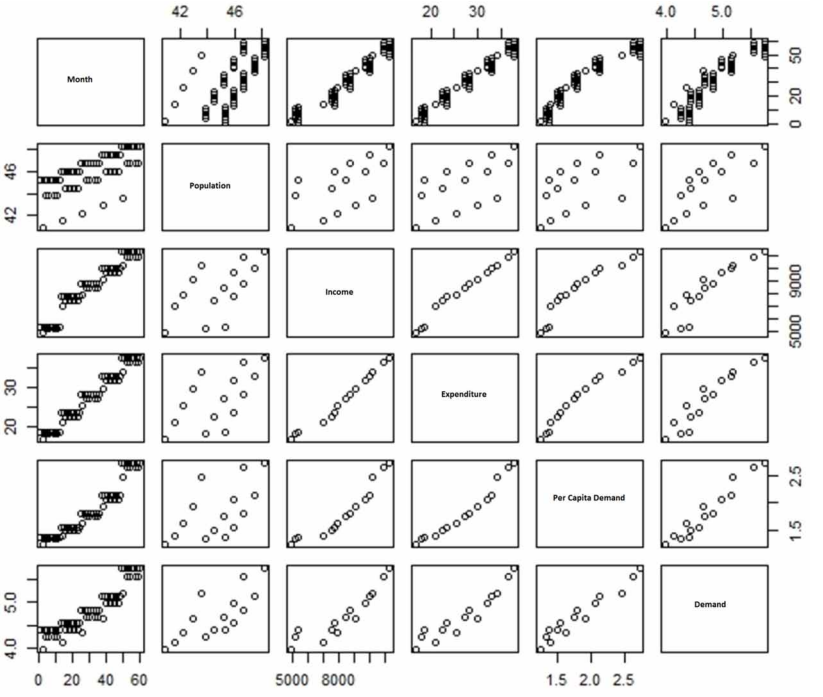
Table 2. Model coefficients

Variable	Estimate	Std. Error	T Value	Significance
Population	2.348e+01	1.268e+00	18.514	***
Income	8.884e-05	2.989e-06	29.725	***
Per capita	-1.199e+01	6.619e-01	-18.116	***
Expenditure	0.159990	0.005299	30.19	***
Multiple R-squared	0.9861			
Adjusted R-squared	0.9859			

Table 3. Comparison of actual and predicted demand values

Actual Demand	Predicted Demand
26.48	26.5
27.5	27.1
28	27.9
28.7	28.4
29.4	29.6
30.8	31.3
32.5	32.3
34.8	34.6
36.9	37.2
39.3	39.9
42.9	42.9
45.3	45.2

Figure 4. Scatter plot – Linear relationship between dependent and independent variables while forecasting the demand for the crop onion



as the response and predictive variables considered are highly correlated and linearly related with high significance. So, the EMLR-DF model developed could be used for effective forecasting of the demand for different food crops. Hence, using the EMLR-DF model the demand for Rice, Wheat, Tomato and Onion for the state of Karnataka have been forecasted for the years 2018 to 2021 and are drawn in Table 6.

Table 4. Correlation between dependent and independent variables

	Population	Income	Expenditure	Per Capita Demand	Demand
population	1.0000000	0.7040245	0.7098395	0.6922396	0.8029122
income	0.7040245	1.0000000	0.9878007	0.9367008	0.9253252
expenditure	0.7098395	0.9878007	1.0000000	0.9709860	0.9577156
Per capita demand	0.6922396	0.9367008	0.9709860	1.000000	0.9853215
demand	0.8029122	0.9253252	0.9577156	0.9853215	1.000000

Table 5. Model coefficients

Variable	Estimate	Std. Error	T Value	Significance
Population	7.101e-02	2.426e-05	2927.430	***
Income	8.719e-07	1.144e-07	7.619	***
Per capita	9.912e-01	3.162e-04	3135.012	***
Expenditure	-1.061e-02	4.944e-05	-214.699	***
Multiple R-squared	0.9942			
Adjusted R-squared	0.9901			

Table 6. Demand forecasted for onion, tomato, rice and wheat using EMLR-DF model year 2018 to 2021

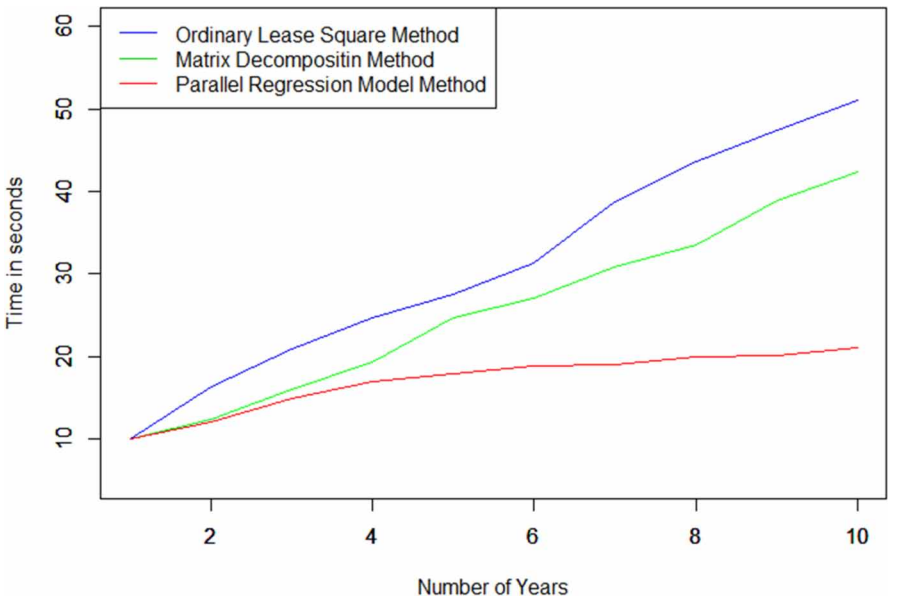
Year	Vegetables		Cereals	
	Onion	Tomato	Rice	Wheat
2018	66.706	62.59364	41.67	10.32
2019	70.026	64.74002	45.08	11.06
2020	73.488	67.34	48.22	11.67
2021	76.95	71.25	52.91	12.33

The time taken by Ordinary Least Square Method, Matrix Decomposition Method and Parallel Multiple Regression Method for computing the coefficient values considering the last ten years datasets are shown in Figure 5. As the EDMLR-DF method performs the computation in parallel using Map-Reduce technique, the average time taken by this method is 14 seconds. Whereas least square method and matrix decomposition methods do not follow parallel computation approach, the average time taken by these methods is 32 seconds and 28 seconds respectively for the same data sets. So, the parallel multiple regression model is more efficient.

7. CONCLUSION

In this work, a truthful demand forecasting model using Multiple Linear Regression [EMLR-DF] is proposed to forecast the demand for various food crops to assist the farmers for growing demand based crops. EMLR-DF model has been developed and evaluated with the sample data sets for the crops

Figure 5. Execution time taken by ordinary least square method, matrix decomposition method and parallel multiple regression method



Rice and Onion considering demand as dependent variable and population, income, per capita demand and expenditure as independent variables. As the response and predictive variables considered are highly correlated, linearly related with high significance of 'p' values and with an R-squared value of 0.98 also from the analysis made with the results of scatter plot and correlation function, it has been concluded that the proposed EMLR-DF is the more accurate model in forecasting the demand for the food crops. The developed model could derive the coefficients in 14 seconds compared to least Square Method, Matrix Decomposition methods which have taken 32 sec and 28 sec respectively. The proposed model has achieved 98% of accuracy in forecasting the demand values for food crops. Hence, this forecasting model could be certainly used to assist the farmers in growing demand based crops and thereby reducing the loss for them also making the customers happy. In future, by making use of this model, a full-fledged system could be developed to assist the farmers.

REFERENCES

Agarwal, M., & Mehra, R. (2014). *Review of Matrix Decomposition Techniques for Signal Processing Applications*. Academic Press.

Demand, *Supply*, *and Market Equilibrium*. (2011). McGraw-Hill.

Demand Analysis and Forecasting. (2017). Retrieved from http://www.oocities.org/ebbins_1/me2.doc

Duncan Waga, K. R. (2014). Environmental Conditions' Big Data Management and Cloud Computing Analytics for Sustainable Agriculture. *World Journal of Computer Application and Technology*, *2*(3), 73–81. doi:10.13189/wjcat.2014.020303

Elsheikh, R., Rashid, A., Shariff, B. M., Amiri, F., Ahmad, N. B., Kumar, S., & Soom, M. et al. (2013). Agriculture Land Suitability Evaluator (ALSE): A decision and planning support tool for tropical and subtropical crops. *Computers and Electronics in Agriculture*, *93*, 98–110. doi:10.1016/j.compag.2013.02.003

Factors of Supply and Demand. (2017). Retrieved from http://www.grainphd.com/wp-content/uploads/2017/07/Supply-and-Demand.pdf

Hans, L. (2017). *Demand Forecasting with Regression Models*. Retrieved from cpdftraining.org/downloads/Levenbach_Causal2017.pdf

Harlalka, R. (2018). *Choosing the Right Machine Learning Algorithm*. Retrieved from https://hackernoon.com/chosing-the-right-machine-learning-algorithm-68126944ce17

Higgins, J. (2005). *Introduction to Multiple Regression Now*. Academic Press.

Ivano, D. (2017). Demand Forecasting. *Global Supply Chain Operation Management Springer Texts in Business Economics*, *10*(11), 301–315. doi:10.1007/978-3-319-24217-0_11

Jiménez, D., Dorado, H., Cock, J., Prager, S. D., & Delerce, S. (2016). *From Observation to Information : Data-Driven Understanding of on Farm Yield Variation*. Academic Press. 10.1371/journal.pone.0150015

João, Z., Mzyece, M., & Kurien, A. (2009). Matrix decomposition methods for the improvement of data mining in telecommunications. *IEEE Vehicular Technology Conference*. doi:10.1109/VETECF.2009.5378904

Kothari, S., Channe, H., Kadam, D., & Professors, A. (2015). Multidisciplinary Model for Smart Agriculture using Internet - of - Things (IoT), Sensors, Cloud - Computing, Mobile - Computing & Big - Data Analysis. *International Journal of Computer Technology & Applications*, *6*, 374–382.

Kumar, R., Singh, M. P., Kumar, P., & Singh, J. P. (2015). Crop Selection Method to Maximize Crop Yield Rate using Machine Learning Technique. *2015 International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy, and Materials (ICSTM)*, 138–145. doi:10.1109/ICSTM.2015.7225403

Kumar, V., & Khan, S. (2017). Importance of weather prediction for sustainable agriculture in Bihar, India. *Archives of Agriculture and Environmental Science*, *2*(2), 105–108.

Mittal, S. (2008). Demand-Supply Trends and Projections of Food in India. Indian Council for Research on International Economics Relations.

Narayana Reddy, M. (2017). GIS-Based Decision Support Systems in Agriculture. NHR.

Nikkilä, R., Seilonen, I., & Koskinen, K. (2010). Software architecture for farm management information systems in precision agriculture. *Computers and Electronics in Agriculture*, *70*(2), 2009–2011. doi:10.1016/j.compag.2009.08.013

Prasad, S., Peddoju, S. K., & Ghosh, D. (2016). AgroMobile : A Cloud-Based Framework for Agriculturists on Mobile Platform AgroMobile : A Cloud-Based Framework for Agriculturists on Mobile Platform. *International Journal of Advanced Science and Technology*, *59*, 41–52. doi:10.14257/ijast.2013.59.04

Simon, G. (2003). Multiple Regression Basics. *Science*, 1–40.

Triplett, J. E. (1976). *Consumer Demand and Characteristics of Consumption Goods*. Academic Press.

Whelan, J., & Forrester, J. W. (1996). *Economic supply & demand*. Academic Press.

Yadav, & Jhalak Rathod. (2015). Big Data Meets Small Sensors in Precision Agriculture. *International Journal of Computer Applications, 201*(2), 1–4.

Balaji Prabhu B. V. is currently perusing Ph.D in the area of Big Data Analytics, in the research centre Information Science and Engineering BMS College of Engineering, Bangalore, Karnataka, India. Areas of interests are Big Data, Data Analytics, Machine Learning, Map Reduce Programing, Demand Supply Management, and Time Series Forecasting.

M. Dakshayini holds Ph.D degree in the area of computer Networks, M.Tech in computer science. She has two decades of experience in teaching field. She has published several research papers in refereed journals. Currently she is working as Professor in the department of Information science and engineering at BMS College of Engineering, Bangalore, India.