Fuzzified input data tuning for agriculture commodities price prediction

Girish Hegde[1], Vishwanath R. Hulipalled[1], Jay B. Simha[2]
[1]School of Computing and Information Technology, REVA University, Bangalore, India
[2]Abiba Systems, CTO, and RACE Labs, REVA University, Bangalore, India

Article Info

Article history:

Received Oct 13, 2023
Revised Feb 22, 2024
Accepted Mar 15, 2024

Keywords:

Agriculture
Anomaly detection
Long short-term memory
Machine learning
Price forecasting

ABSTRACT

The quality of the input data will typically affect the prediction accuracy. Preprocessing of data is commonly referred to as input data tuning. Tuning the input data is critical for projecting commodity prices. Anomalies or outliers are unavoidable in historical price data. To increase prediction and forecasting accuracy, it is necessary to find and correct outliers before training the prediction model. To correct the anomaly and increase prediction accuracy, the fuzzified input data tuning and prediction algorithm proposed in this study. The identified outliers are corrected using the relevant fuzzy set value in this method. With outlier corrected data, we used long short-term memory and seasonal autoregressive integrated moving average to anticipate tomato prices in Karnataka state. The result of the proposed algorithm is compared with the sliding window anomalies correction model, and without disposing of the outliers. The suggested algorithm, with 37.89%, performed better than sliding window with 40.08% and 43.11% mean absolute percentage error, respectively, and without outlier correction. The sensitivity analysis shows that the performance of the model is unaffected by the forecasting horizon. Finally, comparitive analysis peformed with previous research work, and the proposed model performed better.

Corresponding Author:

Girish Hegde
School of Computing and Information Technology, REVA University
Rukmini Knowledge Park, Kattigenahalli, Srinivasa Nagar, Yelahanka, Bangalore 560064, India
Email: girishhegde37@gmail.com

1. INTRODUCTION

Outliers are just unwanted data or noise in the input time series data, which the analyst is not interested in [1]. Anomalies in time series price data affect the performance of prediction and forecasting models, as well as the analytical outcome. The accuracy of the forecasting model is vital when providing information to farmers so that the anticipated price is close to the true value. This is one of the most difficult issues in constructing a prediction or forecasting model. One of the key study areas and important jobs in time series forecasting, data mining, and analysis is input data pre-processing, which includes outlier detection and correction. Techniques for detecting and correcting anomalies have been used in a variety of fields, including stock price predictions, cyber security, and fraud detection.

Time series data can be either univariate or multivariate. It is necessary to remove the noise at that moment in time, or the data must be corrected. For this research, univariate pricing data was utilized. In the case of time series data, outliers are defined as deviations from the expected or standard observation. Outliers in time series data can be generated for a variety of reasons. In the case of stock price data, for example, the outlier could be due to a worldwide recession, extremely high inflation, or other factors. In the case of agricultural commodity prices, it could be due to unseasonable rain, a new government policy, and disease-affected production

Several investigations have been undertaken in recent decades, and numerous approaches for detecting and correcting anomalies have been proposed. Incorrect model selection can have an impact on prediction and forecasting ability. As a result, it is vital to seek a robust model capable of detecting and correcting anomalies in time series data without altering the relationship or performance. The outlier correction model is regarded to be good if it does not impact the time series data's trend or seasonality. Our primary goal is to repair the anomalies found in time series data [2], without impacting the trend or seasonality. Various domains require different types of anomaly detection techniques for different purposes. The authors organized their survey of several outlier detection approaches. There is no universal strategy for dealing with various data kinds. Many outlier detection strategies are available depending on the application domain [3]–[5]. Schmidl *et al.* [6] conducted a detailed study on anomaly detection for time series data and indicated that in the past many techniques are evaluated or developed, but there was no thorough research that analyzes and assesses the various strategies in an organized manner.

Xu *et al.* [7] explored recent advances in outlier detection as well as the benefits and drawbacks of several anomaly detection approaches. Because of equidistant properties, typical neighbour-based approaches will not operate efficiently. The methods based on neighbours are sensitive to the nearest neighbour chosen. The ensemble and subspace approaches perform reasonably well, but selecting the appropriate subspace is difficult. A systematic method of detecting and correcting anomalies is required. Most of the researchers did not adhere to defined procedures while discovering or correcting anomalies in the data set [8]. Sliding window prediction (SWP) based techniques, according to Ranjana *et al.* [9], can be utilized to detect and repair anomalies. SWP techniques work well with less volatile data and are window width dependent. If the data is less volatile, the portrait dataset-based techniques work well. When compared to other raw data smoothing methods, B-spline smoothing approaches perform well. The k-nearest neighbor (KNN) based approaches can detect but not rectify anomalies. If the trend of the original data is not modified, the preprocessing procedure is effective.

Most earlier studies concentrated on identifying anomalies rather than correcting them [10]. Outlier removal is straightforward, but it may introduce bias. Autoregressive [11] and autoregressive with exogenous [12], inputs are popular models for detecting anomalies. Another simple method is to replace the outlier with mean value of the data set [13]. Simple moving average approaches smooth time series data, however, they may change the correct dataset [14]. Another extensively used method is constraint-based correction, however, it cannot handle continuous faults. Price changes will have an impact on both producers and consumers [15]. For anomaly identification, the authors used the random forest classifier and proposed a graph-based model for future work. The use of outlier filters improves forecast accuracy [16]. Outlier detection is critical with pricing data due to price volatility. Previous research on price outlier filtering in forecasting was insufficient. The authors proposed the algorithm tailored for the precise detection of lower outliers in univariate datasets. To effectively separate anomalies from normal values, the method combines sophisticated filtration techniques with transformative processes [17]. The utilization of clustered information concerning the forecast accuracy of the outlier has risen. There is no universal method for detecting and correcting outliers [18]. As the volume of data grows, so does the complexity of computation, and detecting anomalies becomes more difficult. Many alternative outlier identification and correction approaches based on machine learning [19], artificial neural networks [20], and domain-specific [21], have been investigated.

Over the last few decades, many predictive and price forecasting models have been proposed by researchers. The price of tomatoes in India varies between locations due to a lack of cold storage and refrigerated transport [22]. Some algorithms excel at long-term predicting, whereas others excel at short-term forecasting [23]. The autoregressive integrated moving average (ARIMA) model performs better for long-term forecasting, whereas the recurrent neural network model performs better for short-term forecasting. When compared to a single model, hybrid models often improve prediction accuracy. Recent advancements in artificial neural networks and other advanced algorithms such as long short-term memory (LSTM), convolution neural network (CNN) [24], back propagation neural networks (BPNN) [25], [26], and support vector machine (SVM) [27], are particularly promising for increasing prediction accuracy. Smart algorithms will perform better with nonlinear data as well, but one issue with the neural network model is the slow convergence time. Farmers can determine which crop to cultivate in advance if there is a strong price prediction [28]. Predicting commodity prices is a significant issue in agriculture [29]. When projecting the price, the precision of the forecasting model is critical. Advanced models, such as XGBoost and neural networks, will provide greater accuracy.

From earlier research and similar work that forecasting and predicting commodity prices is a challenging endeavor due to market volatility and other contributing variables. Many strategies for detecting anomalies have been proposed, and significantly more research is required to correct anomalies. After correcting the anomalies, most of the models alter the trends and properties of the dataset. Our goal is to repair the anomaly in time series price without compromising the data's relationship and suggested a fuzzy-based outlier correction model and discovered that it improved prediction accuracy. This work's significant contributions are:

- A new fuzzified input data tuning and prediction (FUIDTAP) approach was proposed that corrects the outlier pricing value identified with a fuzzy set value.
- The proposed model has no effect on the original data set's trend or properties and introduces no bias.
- The suggested methodology increased price prediction accuracy, according to a comparative analysis that included outlier correction using the sliding window method, no outlier correction, and imputation using mean and median.
- Forecasting accuracy was compared to one of the earlier studies [22], and it was discovered that the proposed model, FUIDTAP, gave superior accuracy.

The remainder of this paper is organized as follows. Section 2 describes the technique and proposed algorithm in detail. In section 3, described the data set used for this investigation, as well as the experimental results and analyses. Section 4 is dedicated to the conclusion and future work.

2. MATERIALS AND METHOD

2.1. Pre-processing

In the given data set, there may be a few missing price data values. It is critical to deal with missing data before processing the data. The simplest solution is to leave out the missing number, but this creates a gap in the data set. There are numerous imputing methods accessible, such as mean, mode, and nearest neighbours. Linear interpolation is a simple and widely used technique that outperforms the mean method [30]. To fill in the gaps, employed linear interpolation. Let *D* be a price data set of size *n* and $p_{1,}\ p_2, p_3, \dots \dots .., p_n$ are price values. Traverse the entire data set and find the missing price elements and fill in the missing values, as shown in (2).

$$D = \{p_{1,}p_2, p_3, \dots \dots .. p_n\} \tag{1}$$

Let *p* represents the missing price on date d, the interpolation logic is as (2):

$$Price(p) = \frac{d-d_1}{d_0-d_1} * Price(d_0) + \frac{d-d_0}{d_1-d_0} * Price(d_1) \tag{2}$$

Where *d* is the date on which price data is missing, d_0 is the immediately preceding date, d_1 is the immediately following date, *Price*(d_0) and *Price*(d_1) are the known prices before and after the missing price data on the specific date.

2.2. Fuzzification of price data

Song and Chissom [31], [32] proposed the fuzzy time series model or approach. The primary distinction between fuzzy time series and regular time series is that the values in fuzzy time series are fuzzy sets rather than numbers. Complex computations are required for this model. Chen [33] proposed an enhanced and simplified model. Let *D* be the price data set with length *n*. As in (3), let *U* be the universe of discourse and u_1, u_2, u_3……., u_n be the partitions of equal length.

$$U = \{u1,\ u2, u3, \dots \dots \dots ., un\} \tag{3}$$

Let. X_1, X_2 ... X_n are the defined fuzzy sets for *U*. So, each fuzzy set can be defined as:

$$X_1 = \{u1/0,\ u2/1, u3/0, \dots \dots \dots .un/1\ \} \tag{4}$$

$$X_2 = \{u1/1,\ u2/0, u3/1, \dots \dots \dots .un/0\ \} \tag{5}$$

$$X_n = \{u1/0,\ u2/1, u3/0, \dots \dots \dots .un/1\ \} \tag{6}$$

Assign the price value to one of the fuzzy sets specified earlier, as shown in (7).

$$p_1 \to X_1,\ p_2 \to X_1,\ p_3 \to X_2, \dots .,\ p_n \to X_n \tag{7}$$

Identify the fuzzy relation and make the group as in (8) to (11).

$$G_1 = X_1 \rightarrow X_1 , X_1 \rightarrow X_3 , X_1 \rightarrow X_5 \quad (8)$$

$$G_2 = X_2 \rightarrow X_2 , X_2 \rightarrow X_4 \quad (9)$$

$$G_3 = X_3 \rightarrow X_1 , X_3 \rightarrow X_3 \quad (10)$$

$$G_n = X_n \rightarrow X_1 , X_n \rightarrow X_n \quad (11)$$

where $G_1, \dots, G_n$ - different groups. $X_1, \dots, X_n$ - fuzzy relations.

Each pricing value p_i for $0 > i < n$, belongs to one of the groups G_j with a set of fuzzy relations for $0 > j < n$. Once the groups have been formed using the set of fuzzy relations, the average values for each group must be determined, as shown in (12):

$$Avg(G_k) = (\sum_{i=1}^{n} Avg(u_i))/n \quad (12)$$

where G_k is the specific group with the average value to be found. k *is* $0 > k <= n$, u_i is one of the partitions with the universe of discourse *U*. For example, as shown in (13):

$$G_2 = X_2 \rightarrow X_2 , X_2 \rightarrow X_4. \textit{Then}$$

$$Avg(G_2) = (Avg(u_2) + Avg(u_4))/2 \quad (13)$$

2.3. Anomaly detection and correction

The isolation forest is one of the popular decisions tree-based method was used to find anomalies or outliers. A couple of outliers were found and represented as in (14), using the data set TD of size n.

$$D = \{p_{1,}p_2, p_{o3}, \dots p_{ok} \dots p_{ol} \cdot\cdot p_n\} \quad (14)$$

where $p_{1,}p_2, \dots, p_n$ are price values. p_{o3}, p_{ok} *and* p_{ol} are outliers.

The next step is to determine which fuzzy set and group each outlier belongs to, and then replace the actual price value with the average value obtained using fuzzy groups $G_1, G_2, \dots, G_n$.
Let $p_{o3} \rightarrow G_3$ $p_{ok} \rightarrow G_k$ and $p_{ol} \rightarrow G_l$ then

$$pa_{o3} = Avg(G_3) \quad (15)$$

$$pa_{ok} = Avg(G_k) \quad (16)$$

$$pa_{ol} = Avg(G_l) \quad (17)$$

The final price data set *D* will look like this after anomaly identification and replacement with the fuzzy set value and can be used to predict and forecast commodity prices.

$$D = \{p_{1,}p_2, pa_{o3}, \dots pa_{ok} \dots pa_{ol} \cdot\cdot p_n\} \quad (18)$$

Where pa_{o3}, pa_{ok} *and* pa_{ol} are the replaced fuzzy set values for the corresponding price data.

2.4. Proposed model

Figure 1, depicts the data flow of the proposed model. The price data presented here includes historical commodity price data, such as monthly, weekly, or daily commodity prices. For this study, collected monthly price data for tomatoes and the state of Karnataka in India from January 2011 to December 2020.

The subsequent step is to preprocess the price data. The performance of forecasting algorithms is influenced by data quality. It is critical to fine-tune the input data before beginning the prediction procedure. First, determine whether any particular month's pricing data is missing. If information is absent, it is filled in utilizing linear interpolation logic. After adding the missing pricing data, divide the data set into training and testing sets. Identifying outliers is crucial to increase the quality of the training data even further. Because of a certain reason, the price may fluctuate substantially. The pricing data set can show a quick surge or decline. To increase the data set's quality, the fuzzification approach was applied. Then it is required to partition the universe of discourse into equal portions during the fuzzification process, as stated in (3).

After creating the partitions, define the fuzzy set and assign each price data point from the training data set to one of the fuzzy sets. This is simply setting the linguistic value for the numeric price value, such as $100 \rightarrow X_1$, $753 \rightarrow X_2$ and so on. Once the fuzzy set has been constructed, identify and group the fuzzy relations. A set of relations for the same value may exist, and it is required to identify and classify them as illustrated in (19). A group can be composed of several relationships for a given value, and each of them has its own partitions in its discourse universe, such as:

$$G_2 = X_2 \rightarrow X_2 , X_2 \rightarrow X_4 \tag{19}$$

$$G_3 = X_3 \rightarrow X_3 , X_3 \rightarrow X_2 , X_3 \rightarrow X_4 \tag{20}$$

$$G_2 = \{u_2 , u_4\} \tag{21}$$

$$G_3 = \{u_2 , u_3 , u_4\} \tag{22}$$

Calculate the mean value for each identified group for each detected outlier, as illustrated in (13). As an example:

$$Avg(G_3) = (Avg(u_2) + Avg(u_3) + Avg(u_4))/3 \tag{23}$$

Replace the outlier value with the average value of the related fuzzy group in the final pre-processing step. Pre-processing and fuzzified outlier values result in tuned data in the training set. After tuning the input training data set, utilize it to train the predictive machine learning algorithm. Improved prediction performance was obtained by modifying the data set.

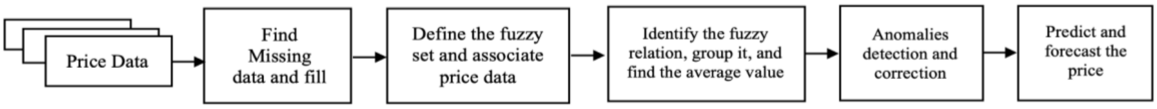
Price Data
Find Missing data and fill
Define the fuzzy set and associate price data
Identify the fuzzy relation, group it, and find the average value
Anomalies detection and correction
Predict and forecast the price

Figure 1. Data fuzzification, anomaly detection, correction, and prediction process

2.5. Algorithm

The proposed Algorithm 1 describes FUIDTAP in detail. In the phase 1 it will impute the missing data point using linear interpolation method. The second phase is the fuzzification of the price data, detect the outlier and correct the outliers in the data set. In the phase 3 the data with outlier corrected was used to train the prediction model and forecate the price. The performance of the prediction algorithm was calculated with mean absolute percentage error (MAPE).

```
Algorithm 1: Fuzzified input data tuning and prediction
Input: Price data ($D$)
Output: Predicted price ($P$)
begin
    Phase – 1: Find the missing price data and fill
        for each $i$ in $D$
            if ($D[i]$) is empty then
                fill the data with linear interpolation
        end for
    Phase – 2: Fuzzification
      $n \leftarrow$ length($D$)
      #Partition  the universe of discourse
            $U \leftarrow \{u_1,u_2,u_3, \ldots\ldots, u_n\}$
      #Define the fuzzy sets , $X_1, X_2, X_3, \ldots,X_n$
      #Associate each data with one of the fuzzy set
        for each $value$ in $D$
          if ($value$) belongs to $u_k$  then
            associate $value$ with $X_k$  where  $0>k<n$
        end for
      #Find the fuzzy relation and group it
      #Find the mean for each group $G$
        for each $value$ in $U$
          mean($G_i$) ← ((mean($u_k$)+ mean($u_l$) + … + mean($u_r$))/s)
            where  0>$k$,$l$,$r$<$n$ , $s$ = number of fuzzy relation.
        end for
      #Detect the outlier in the data set
      #Replace the outlier with corresponding fuzzy group value
        for each $i$ ←1 to $n$
          if ($D$[$i$] is outlier  and belongs to $G_i$) then
            $D$[$i$]  ← mean($G_i$)
        end for
    Phase – 3: Train the prediction model with tuned data and  forecast
        training_set ←  80% of $D$
        testing_set ←  20% of $D$
        fit ← model(training_set)
        predict ← fit.predict(testing_set)
      #Find the performance
        mape = (mean (testing_set.Price – predict.Price )/ testing_set.Price)* 100
        Forecast the price ($P$)
end
```

3. RESULTS AND DISCUSSION

3.1. Data set

Tomato is a vegetable or agricultural item that is consumed by the majority of Indians all year. The tomato pricing data from previous years was used to conduct this research. AGMARKNET [34] is used to collect the pricing data. To evaluate the algorithm's efficacy, we obtained monthly tomato pricing data for Karnataka, India's largest producing state. The price for 100 kg is in Indian Rupees (INR). We used monthly average price data for tomatoes from January 2011 to December 2020 that included 10 years of historical average prices for our analysis. The monthly average tomato price is displayed in Figure 2.

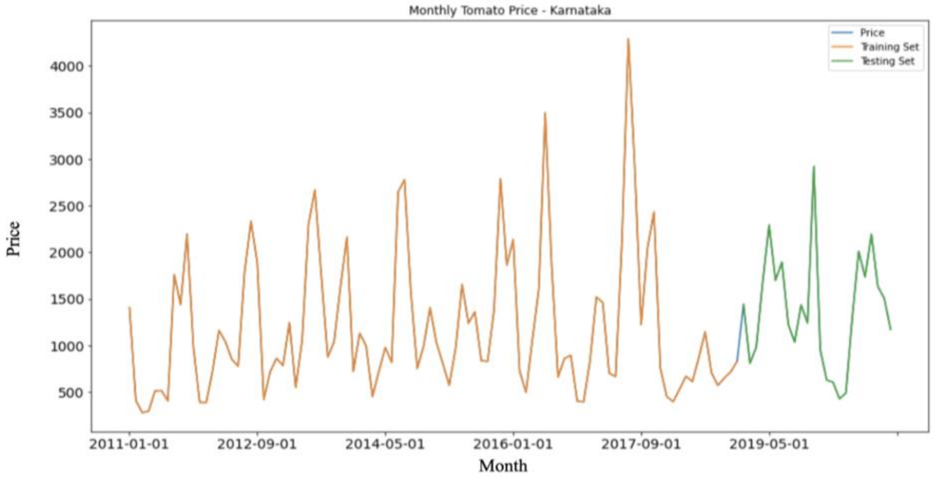

Figure 2. Monthly average tomato price

The aim of this was to demonstrate the need for input data tuning as well as how we might enhance prediction accuracy by fuzzifying outliers. The LSTM and seasonal autoregressive integrated moving average (SARIMA) algorithms are used to predict commodity prices for this reason. When compared to outlier correction with the Sliding Window model and without outlier correction, the adoption of a suggested algorithm, FUIDTAP, enhances the accuracy of the projected values.

3.2. Long short-term memory

LSTM is one of the widely used deep learning algorithms. This is a kind of recurrent neural network and gains the ability to learn long-term dependencies. LSTM can be used for univariate or multivariate time series prediction problems. Unlike the popular parametric models like ARIMA and SARIMA [35], there is no need for stationary data and forecasting performance is better compared to traditional models [36]. The LSTM layer consists of a set of recurrently connected memory blocks. LSTM has feedback connections and can keep the information for a long time. Even in the case of noise, LSTM can learn time intervals more than 1000 steps, with the help of efficient gradient-based algorithms. The LSTM memory block contains one or more memory cells and shares the same gates to reduce the parameters. The values are retained or discarded based on the state of the gates (1 or 0). The Figure 3 shows the LSTM cell. For the given input data $x = x_{1,} \dots., x_t$ and the output calculated by network $y = y_{1,} \dots., y_t$ then the activation function can be represented as:

$$i_i = \sigma(W_{xi}\, x_t + W_{hi}\ h_{t-1} + W_{ci}\, c_{t-1} + b_i) \tag{24}$$

$$f_i = \sigma(W_{xf}\, x_t + W_{hf}\ h_{t-1} + W_{cf}\, c_{t-1} + b_f) \tag{25}$$

$$c_t = f_t\, c_{t-1} + i_i\, tanh\,(W_{xc}\ x_t + W_{hc}\ h_{t-1} + b_c) \tag{26}$$

$$o_t = \sigma(W_{xo}\, x_t + W_{ho}\ h_{t-1} + W_{co}\, c_{t-1} + b_o) \tag{27}$$

$$i_t = o_t tanh\,(c_t) \tag{28}$$

where $\boldsymbol{\sigma}$ is logistic sigmoid function, $\boldsymbol{i}$ is input gate, $\boldsymbol{f}$ is forget gate, $\boldsymbol{o}$ is output gate, $\boldsymbol{c}$ is activation vector, $\boldsymbol{W}$ is weight matrices, and $\boldsymbol{tanh}$ is activation function

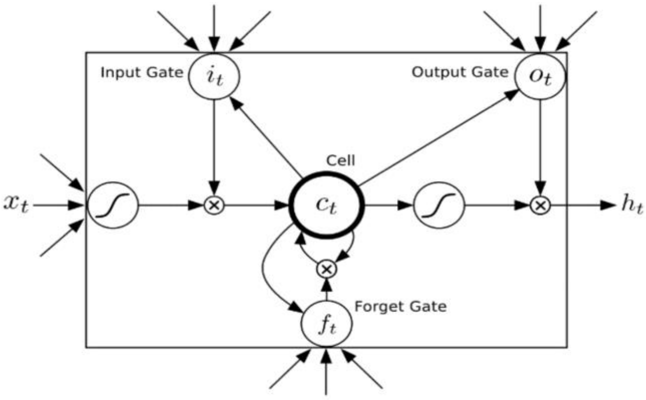

Figure 3. LSTM Cell

3.3. Sliding window model

Outlier detection and correction are commonly performed using the sliding window model [37], [38]. The sliding window model requires the size of the window to be determined. The entire data set is divided into several windows of a specified length after the window size is defined. After detecting the outlier, it is required to find the window where the outlier data is located. Then replace the outlier with the window mean value. Let data set *D* be divided into N number of widows, *W*, of size n.

$$D = W_1, W_2, \dots\dots, W_N \tag{29}$$

$$W_i = p_1, p_2, \dots\dots, p_n \tag{30}$$

Where 0 > *i* < *n*, $\boldsymbol{p_1}, \boldsymbol{p_2}, \dots\dots, \boldsymbol{p_n}$ is price data. Let $\boldsymbol{p_t}$ be an outlier belonging to a window $\boldsymbol{W_i}$, then $\boldsymbol{p_t = Avg(W_i)}$. Repeat the same for all outliers detected within the given dataset. Figure 4 shows a simple sliding window of size 3.

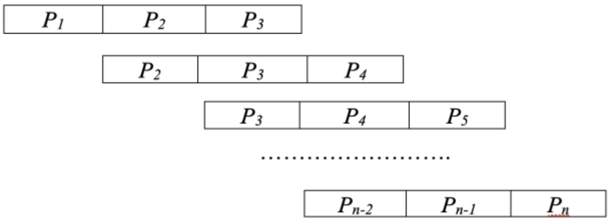

Figure 4. Sliding window

3.4. Data imputation

The data imputation is one of the commonly used outlier correction methods. The mean, often known as the arithmetic mean, is the average of all numbers. Determine the mean value of the price data and use it to replace the anomaly in the outlier correction using mean technique. The middle value in a set of numbers is called the median. The price value of the discovered outlier is replaced with the median of the price data set when using the outlier correction by median approach. The number that appears the most frequently among a range of numbers is the mode. Imputation by mode refers to the process of substituting the mode value for the outlier value in a price data collection.

3.5. Model performance evaluation

The statistical parameters such as MAPE, root mean square error (RMSE), and mean absolute error (MAE) were used to measure the model's performance. MAPE is a measure of how accurate a forecasting system is. It measures this accuracy as a percentage and can be calculated as the average absolute percent error for each time the actual value minus predicted values is divided by the actual value.

$$MAPE = \frac{100}{n}\sum_{t=1}^{n} |A_t - \frac{F_t}{A_t}| \tag{31}$$

The RMSE provides the standard deviation value. The RMSE is a square root of MSE. This helps us to evaluate the usefulness and accuracy of the prediction model.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)} \tag{32}$$

The MAE provides the average residue in the given dataset. This is the mean of the absolute difference between the actual and expected values in the data set.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |\hat{y}_i - y_i| \tag{33}$$

3.5. Prediction results

The proposed model was used to determine the results with the historical price of commodity tomatoes in Karnataka state. Figure 5 provides the details of the results obtained with the different outlier correction models. LSTM and SARIMA was used for price prediction. The experimental results clearly indicate that the proposed FUIDTAP performed better than those without outlier correction and the Sliding window model. In the case of the sliding window model, the outlier price data is replaced with the average value of the window of specified size (the defined window size was 3). The performance comparison for the various models is shown graphically in Figure 5.

The proposed model and commodity tomato resulted in a 37.89% MAPE compared to 43.11% with no outlier correction, resulting in a 5.22% improvement in prediction performance. Compared to the sliding window mechanism of 40.08% MAPE, a performance improvement of 2.19% achieved. Compared to outlier correction with mean and meadian methods, accuracy improvemt got was 1% and 7% respectively. With the proposed model each outlier data was considered independently and replaced by considering the relation of that datapoint across the different fuzzyset. Unlike mean/mode/median method outlier was not replaced with the single value or average value of specific window in case of sliding widow. When compared to other widely used methods, the prediction performance increased when the outlier value was substituted with data that was extremely close to the real value.

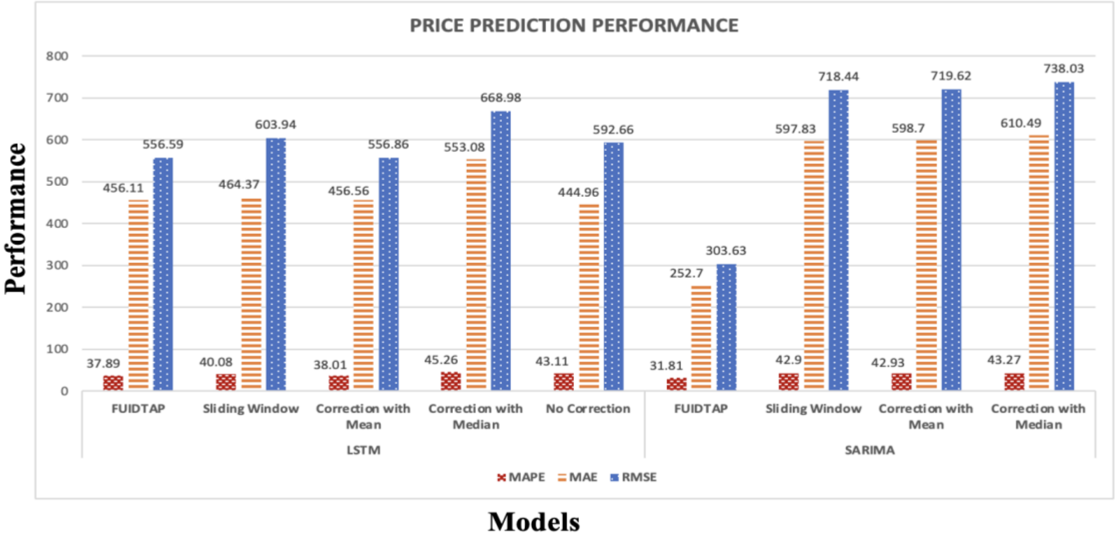

Figure 5. Price prediction performance

3.6. Performance comparison with previous research work

The result of the proposed model was compared to one of the previous studies [22]. The author used SARIMA to project the price of tomatoes. The historical monthly price between January 2006 and December 2016 was used. The same dataset was used for Karnataka and the performance comparison is presented in Table 1. The experimental results show that the accuracy has increased with the proposed model, compared to previous work (SARIMA). The comparison result is graphically displayed as shown in Figure 6.

Table 1. Performance comparison

	FUIDTAP	SARIMA
MAPE	29.88	36.03
MAE	245.29	285.5
RMSE	497.69	413.8

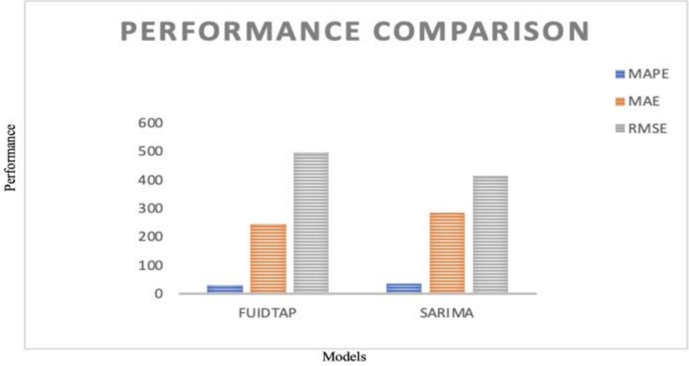

Figure 6. Comparison grapic of performance

4. CONCLUSION

In this study, a novel FUIDTAP algorithm was proposed and that will help to identify the outlier in the input data and replace the same with a fuzzified value. The experimental results show that by using these adjusted data, the prediction performance will increase, and we will be able to get more accurate predictions and forecast prices. The proposed algorithm showed a 2.19% improvement in accuracy compared to the sliding window method, and a 5.22% improvement in accuracy compared to no outlier correction. The comparative study shows that with the proposed FUIDTAP, the forecast accuracy increased by 6% compared to SARIMA without anomaly correction. The sensitivity analysis results suggest that the model's performance is not affected by forecasting horizons. The advantages of the proposed model are, improved prediction accuracy, there is no bias and a reduction in the data set size and it will not affect the trend. In the future, more outliers can be detected by utilizing other outlier detection mechanisms like one-class SVM, minimum covariance determinant, and local outlier factor. Impute the data using methods like adaptive multipath liner interpolation, can be explored. Improve prediction performance by incorporating advanced deep learning algorithms that consider multiple parameters such as production and weather conditions.

REFERENCES

[1] C. C. Aggarwal, *Outlier Analysis*, Cham: Springer Publishing, 2017, doi: 10.1007/978-3-319-47578-3.

[2] A. Puder, M. Zink, L. Seidel, and E. Sax, “Hybrid anomaly detection in time series by combining kalman filters and machine learning models,” *Sensors*, no. 9, 2024, doi: 10.3390/s24092895.

[3] R. Kiani, W. Jin, and V. S. Sheng, “Survey on extreme learning machines for outlier detection,” *Machine Learning*, 2024, doi: 10.1007/s10994-023-06375-0.

[4] S. Baccari, M. Hadded, H. Ghazzai, H. Touati, and M. Elhadef, “Anomaly detection in connected and autonomous vehicles: a survey, analysis, and research challenges,” *IEEE Access*, vol. 12, pp. 19250–19276, 2024, doi: 10.1109/ACCESS.2024.3361829.

[5] M. Olteanu, F. Rossi, and F. Yger, “Meta-survey on outlier and anomaly detection,” *Neurocomputing*, vol. 555, 2023, doi: 10.1016/j.neucom.2023.126634.

[6] S. Schmidl, P. Wenig, T. Papenbrock, “Anomaly detection in time series: a comprehensive evaluation,” *Proceedings of the VLDB Endowment*, vol. 15, no. 9, pp.1779-1797, 2022, doi: 10.14778/3538598.3538602.

[7] X. Xu, H. Liu, and M. Yao, “Recent progress of anomaly detection,” *Complexity*, vol. 2019, pp. 1–11, 2019, doi: 10.1155/2019/2686378.

[8] E. J. Jamshidi, Y. Yusup, J. S. Kayode, and M. A. Kamaruddin, “Detecting outliers in a univariate time series dataset using unsupervised combined statistical methods: A case study on surface water temperature,” *Ecological Informatics*, vol. 69, 2022, doi: 10.1016/j.ecoinf.2022.101672.

[9] K. G. Ranjan, B. R. Prusty, and D. Jena, “Review of preprocessing methods for univariate volatile time-series in power system applications,” *Electric Power Systems Research*, vol. 191, 2021, doi: 10.1016/j.epsr.2020.106885.

[10] A. Zhang, S. Song, J. Wang, and P. S. Yu, “Time series data cleaning: from anomaly detection to anomaly repairing,” *Proceedings of the VLDB Endowment*, vol. 10, no. 10, pp. 1046–1057, 2017, doi: 10.14778/3115404.3115410.

[11] A. E. Bilecen, A. Ozalp, M. S. Yavuz, and H. Ozkan, “Video anomaly detection with autoregressive modeling of covariance features,” *Signal, Image and Video Process*, vol. 16, pp. 1027–1034, 2022, doi: 10.1007/s11760-021-02049-3.

[12] M. Bai, J. Liu, J. Chai, X. Zhao, and D. Yu, “Anomaly detection of gas turbines based on normal pattern extraction,” *Applied Thermal Engineering*, vol. 166, 2020, doi: 10.1016/j.applthermaleng.2019.114664.

[13] C. Yu, J. Tan, Y. Cheng, and X. Mi, “Data analysis and preprocessing techniques for air quality prediction: a survey,” *Stochastic Environmental Research and Risk Assessment*, pp. 1-23, 2024, doi: 10.1007/s00477-024-02693-4.

[14] A. Ermakov and L. Suchkova, “Pre-processing of observation data of intelligent agents using real-time causal filters,” *AIP Conference Proceedings*, vol. 2948, no. 1, 2023. doi: 10.1063/5.0165237.

[15] L. Madaan, A. Sharma, P. Khandelwal, S. Goel, P. Singla, and A. Seth, “Price forecasting & anomaly detection for agricultural commodities in India,” *COMPASS 2019 - Proceedings of the 2019 Conference on Computing and Sustainable Societies*. ACM, pp. 52–64, 2019. doi: 10.1145/3314344.3332488.

[16] D. O. Afanasyev and E. A. Fedorova, “On the impact of outlier filtering on the electricity price forecasting accuracy,” *Applied Energy*, vol. 236, pp. 196–210, 2019, doi: 10.1016/j.apenergy.2018.11.076.

[17] E. M Limam, I. Bellami, and A. Tmi, “Univariate Outlier detection: precision-driven algorithm for single-cluster scenarios,” *Preprints*, 2024, doi: 10.20944/preprints202404.2008.v1.

[18] S. Thudumu, P. Branch, J. Jin, and J. J. Singh, “A comprehensive survey of anomaly detection techniques for high dimensional big data,” *Journal of Big Data*, vol. 7, no. 1, 2020, doi: 10.1186/s40537-020-00320-x.

[19] G. Muruti, F. A. Rahim, and Z.-A. B. Ibrahim, “A survey on anomalies detection techniques and measurement methods,” *2018 IEEE Conference on Application, Information and Network Security (AINS)*, pp. 81–86, 2019. doi: 10.1109/ains.2018.8631436.

[20] B. Lindemann, B. Maschler, N. Sahlab, and M. Weyrich, “A survey on anomaly detection for technical systems using LSTM networks,” *Computers in Industry*, vol. 131, 2021, doi: 10.1016/j.compind.2021.103498.

[21] T. H. A. Musa and A. Bouras, “Anomaly detection: a survey,” *Lecture Notes in Networks and Systems*, vol. 217, pp. 391–401, 2022, doi: 10.1007/978-981-16-2102-4_36.

[22] A. A. Reddy, “Price forecasting of tomatoes,” *International Journal of Vegetable Science*, vol. 25, no. 2, pp. 176–184, 2019, doi: 10.1080/19315260.2018.1495674.

[23] Y. Weng, X. Wang, J. Hua, H. Wang, M. Kang, and F. Y. Wang, “Forecasting horticultural products price using ARIMA model and neural network based on a large-scale data set collected by web crawler,” *IEEE Transactions on Computational Social Systems*, vol. 6, no. 3, pp. 547–553, 2019, doi: 10.1109/TCSS.2019.2914499.

[24] J. Q. Li, “Research on fruit price forecasting and fluctuation early-warning,” Huazhong Agricultural University, 2015.

[25] Y. Yu, H. Zhou, and J. Fu, “Research on agricultural product price forecasting model based on improved BP neural network,” *Journal of Ambient Intelligence and Humanized Computing*, 2018, doi: 10.1007/s12652-018-1008-8.

[26] K. G. Preetha *et al.*, “Price forecasting on a large scale data set using time series and neural network models,” *KSII Transactions on Internet and Information Systems*, vol. 16, no. 12, 2022, doi: 10.3837/tiis.2022.12.008.

[27] J. Xie, “Research on price forecasting of gannan navel based on BP neural network,” Huazhong Agricultural University, 2017.

[28] B. Wang *et al.*, “Research on hybrid model of garlic short-term price forecasting based on big data,” *Computers, Materials and Continua*, vol. 57, no. 2, pp. 283–296, 2018, doi: 10.32604/cmc.2018.03791.

[29] R. K. Paul *et al.*, “Machine learning techniques for forecasting agricultural prices: a case of brinjal in Odisha, India,” *PLoS ONE*, vol. 17, no. 7 July, pp. e0270553–e0270553, Jul. 2022, doi: 10.1371/journal.pone.0270553.

[30] N. M. Noor, M. M. A. B. Abdullah, A. S. Yahaya, and N. A. Ramli, “Comparison of linear interpolation method and mean method to replace the missing values in environmental data set,” *Materials Science Forum*, vol. 803, pp. 278–281, 2015, doi: 10.4028/www.scientific.net/MSF.803.278.

[31] Q. Song and B. S. Chissom, “Forecasting enrollments with fuzzy time series - Part I,” *Fuzzy Sets and Systems*, vol. 54, no. 1, pp. 1–9, 1993, doi: 10.1016/0165-0114(93)90355-L.

[32] Q. Song and B. S. Chissom, “Fuzzy time series and its models,” *Fuzzy Sets and Systems*, vol. 54, no. 3, pp. 269–277, 1993, doi: 10.1016/0165-0114(93)90372-O.

[33] S. M. Chen, “Forecasting enrollments based on fuzzy time series,” *Fuzzy Sets and Systems*, vol. 81, no. 3, pp. 311–319, 1996, doi: 10.1016/0165-0114(95)00220-0.

[34] “AGMARKNET,” *Agmarknet.* Accessed: Mar. 02, 2023. [Online]. Available: https://agmarknet.gov.in.

[35] C. Nwokike, and E. O. Okereke, “Comparison of the performance of the SANN, SARIMA and ARIMA models for forecasting quarterly GDP of Nigeri,” *Asian Research Journal of Mathematics*, vol. 17, no. 3, pp. 1-20, 2021, doi: 10.9734/arjom/2021/v17i330280.

[36] S. S. -Namini, N. Tavakoli, and A. S. Namin, “A comparison of ARIMA and LSTM in forecasting time series,” *17th IEEE International Conference on Machine Learning and Applications, ICMLA*, IEEE, pp. 1394–1401, 2018. doi: 10.1109/ICMLA.2018.00227.

[37] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.

[38] K. G. Ranjan, D. S. Tripathy, B. R. Prusty, and D. Jena, “An improved sliding window prediction-based outlier detection and correction for volatile time-series,” *International Journal of Numerical Modelling: Electronic Networks, Devices and Fields*, vol. 34, no. 1, 2021, doi: 10.1002/jnm.2816.

BIOGRAPHIES OF AUTHORS

Girish Hegde is a staff software engineer at Service Now and a research scholar in the School of Computing & IT, REVA University, Bangalore, Karnataka, India. He completed B.E. in Computer Science and Engineering and M.E. in Computer Science and Engineering. His area of interest includes AI and machine learning, cloud computing, and data analytics. He can be contacted at email: girishhegde37@gmail.com.

Vishwanath R. Hulipalled is a professor in the School of Computing & IT, REVA University, Bangalore, Karnataka, India. He completed B.E., M.E. and Ph.D. in Computer Science and Engineering. His area of interest includes machine learning, natural language processing, data analytics, and time series mining. He has more than 24 years of academic experience and research. He authored more than 50 research articles in reputed journals and conference proceedings. He can be contacted at email: vishwanth.rh@reva.edu.in.

Jay B. Simha is the CTO of ABIBA Systems and Chief Mentor at RACE Labs, REVA University. He completed his B.E. (Mech), M.Tech. (Mech), and M.Phil. (CS) and Ph.D. (AI). His area of interest includes fuzzy logic, soft computing, machine learning, deep learning, and applications. He has more than 20 years of industrial experience and 4 years of academic experience. He has authored/co-authored more than 50 journal/conference publications. He can be contacted at email: jay.b.simha@reva.edu.in; jay.b.simha@abibasystems.com.