# A Forecasting Method Based On K-Means Clustering and First Order Fuzzy Time Series

**S. Imalin[1], V. Anithakumari[2], V.M. Arul Flower Mary[3]**

[1]Research Scholar, Register No.21113042092011 Department of Mathematics, Holy Cross College (Autonomous), Nagercoil – 629004, Email: imalinjas25@gmail.com
[2]Assistant Professor, Department of Statistics, Government Arts and Science College, Nagercoil-629004, Email: anithakumari_v@yahoo.co.in
[3]Associate Professor, Department of Mathematics, Holy Cross College (Autonomous), Nagercoil – 629004, Email: arulflowermary@gmail.com
Affiliated to Manonmaniam Sundranar University, Abishekapatti, Tirunelveli-627012, Tamilnadu, India.

Received: 13.04.2024 Revised : 16.05.2024 Accepted: 23.05.2024

**ABSTRACT**
Clustering is the method of partitioning or grouping a given set of designs into several clusters. Forecasting accuracy is one of the most favourable critical issues in Fuzzy Time Series models. In the past few decades, a number of forecasting models built on fuzzy time series principles have been place out. These models have been frequently used to solve many different types of problems, particularly those involving predicting issues when the event data are linguistic values. The time series forecasting is based on the historical data of 40 years. This study examines introduce a novel fuzzy time series forecasting model that takes historical data as the universe of discourse, cluster the universe of discourse using the K-means clustering approach, and then divides the clusters into intervals. K-mean clustering algorithm was applied by using the IDL software to find the centroid values. The suggested approach is used to forecast data on coal production. At the end, we compared the forecasting values of K-means clustering method and the arithmetic method.

**Keywords**: fuzzy time series; IDL software; K-means; forecasting; coal production;

## 1.INTRODUCTION

It is challenging to understate India's future energy needs. India is the third-largest energy consumer in the world. It is one of the economies with the greatest growth rates and has the second-largest population in the world. The demand for energy in India is likely to rise in the future due to the twin pressures of economic and population growth. Although the country has made great strides in decreasing energy poverty, 168 million people still lived without access to power in 2017 and reliability is still a problem. India's power market also faces a number of other difficulties, which have an impact on the coal industry.

In order to deal with enrolment forecasting, Wang [8] investigated using a high-order time variant fuzzy time series approach. In order to enhance the forecast of enrollments, Huarng [3] presented a heuristic approach for fuzzy time series. Jilani [4] defined the fuzzy sets using a triangle function. In this study, we introduce a novel forecasting approach based on k-means clustering of coal production data. We first choose the coal production data as the universe of discourse. Then, to cluster the data into various-sized intervals, we offer the k-mean clustering approach. We might suggest a novel approach to forecast the data for coal production based on the newly found intervals.

### Fuzzy Time series

Fuzzy with an associated membership function and a fuzzy variable, time series is thought to be a method for solving the fuzzy time series model provided by Song and Chissom (1993) is outlined in the following phases. A time series is a succession of events of observation, which is carried out at regular intervals, is the basis for investigating actual processes in fields such as economics, meteorology, and the natural sciences, among others.

### Fuzzy Cluster

Fuzzy clustering is also referred to as soft clustering or soft k-means and it is a form of clustering in which each data point can belong to more than one cluster.

### K-means Clustering

K-Means Clustering is an Unsupervised Machine Learning algorithm, which divides the unlabelled dataset into many clusters.This is how the algorithm operates:

1. To start, we initialise k means or cluster centroids at random.
2. Each item is categorised according to the closest mean, and the coordinates of that mean which are the average of the items categorised in that cluster thus far are updated.
3. After a specified number of repetitions, we repeat the process, and the result is our clusters.

## 2. METHADOLOGY

### A. Forecasted Method I: A New Approach to forecasting Fuzzy Time Series using K means Clustering

This section outlines the proposed approach's step-by-step process for fuzzy time series forecasting using historical time series data. The suggested method is then used to forecast data related to coal production.

**Step 1**:

The coal production data was divided into 4 clusters using the K-means clustering algorithm.

**Step 2**:

Calculate the cluster centercenter$_m$ shown in table 1 of each cluster m cluster as follows:

$$\mathbf{center_m = \frac{\sum_{j=1}^{n} d_j}{n}}$$

**Step 3:**

Adjust the clusters into intervals according to the follow rules. Assume that center$_m$ and center$_{m+1}$ are adjacent cluster centers, then the upper bound UBound$_m$ and the lower bound LBound$_{m+1}$ of cluster$_{m+1}$ shown in table 1 can be calculated as follows:

$$\mathrm{UBound_m} = \frac{\mathrm{center_m} + \mathrm{center_{m+1}}}{2}$$

LBound$_{m+1}$ = UBound$_m$

where m = 1, 2, ..., k -1. Because there is no too early cluster before the first cluster and there is no next cluster after the last cluster, the lower bound LlBound$_1$of the first cluster and the upper bound UBound$_k$of the last cluster can be calculated as follows:

UBound$_k$ = center$_k$+ (center$_k$ − LBound $_k$)

LBound $_1$ = center$_1$ – (UBound$_k$ − center$_1$)

**Step 4:**

Define each fuzzy set X$_i$based on the intervals and the historical enrollments shown in table 1, where fuzzy setX$_i$denotes a linguistic value of the enrolment's represented by a fuzzy set. As in [4], we use a triangular function to define the fuzzy sets X$_i$.

**Step 5**:

Defuzzify the fuzzy data using the forecasting formula

$$t_j = \begin{cases} \dfrac{1.5}{\frac{1}{a_1} + \frac{0.5}{a_2}}, & \text{if, } j = 1 \\ \dfrac{2}{\frac{0.5}{a_{j-1}} + \frac{1}{a_j} + \frac{0.5}{a_{j+1}}}, & \text{if, } 2 \le j \le n-1 \\ \dfrac{1.5}{\frac{0.5}{a_{n-1}} + \frac{1}{a_n}}, & \text{if, } j = n \end{cases}$$

K-mean clustering algorithm was applied by using the IDL software to find the centroid values. If the centroid values are same then it stops the iteration. It helps to make cluster for classification.

**Table 1.** The intervals generation process from the clusters of the coal production

| Cluster | Data | Center | L bound | U bound | Middle value |
|---|---|---|---|---|---|
| 0 | {113.9,124.2,130.5,138.2,147.4,154.2,165.8,179.7,194.6,200.9,211.7} | 160 | 98.746 | 221.254 | 160 |
| 1 | {229.3,238.3,246.0,253.8,270.1,285.7,295.9,292.3,300,309.6,327.8,341.} | 282.508 | 221.254 | 352.22 | 286.737 |
| 2 | {361.3,382.6,407,430.8,457.1,492.8} | 421.933 | 352.22 | 518.60 | 435.41 |
| 3 | {532,532.7,540,556.4,565.8,609.2 | 615.273 | 518.60 | 741.946 | 615.273 |

| | | | | | |
|---|---|---|---|---|---|
| | ,639.2,657.8,675.4,728.7,730.8} | | | | |

Above table (1) shows the center of cluster, cluster lower bound, cluster upper bound and middle values respectively. Where $a_{j-1}, a_j, a_{j+1}$ are the midpoints of the fuzzy intervals$X_{j-1}, X_j, X_{j+1}$ respectively. $t_j$yields athe predicted values.

## 3. RESULT AND DISCUSSION

**Table 2.** Forecasting values of the coal production data

| Year | Coal production data | Fuzzy Set | Forecast | Year | Coal production data | Fuzzy set | Forecast |
|---|---|---|---|---|---|---|---|
| 1981 | 113.9 | $X_0$ | 187.6464 | 2001 | 309.6 | $X_1$ | 257.7034 |
| 1982 | 124.2 | $X_0$ | 187.6464 | 2002 | 327.8 | $X_1$ | 257.7034 |
| 1983 | 130.5 | $X_0$ | 187.6464 | 2003 | 341.3 | $X_1$ | 257.7034 |
| 1984 | 138.2 | $X_0$ | 187.6464 | 2004 | 361.3 | $X_2$ | 412.1084 |
| 1985 | 147.4 | $X_0$ | 187.6464 | 2005 | 382.6 | $X_2$ | 412.1084 |
| 1986 | 154.2 | $X_0$ | 187.6464 | 2006 | 407.0 | $X_2$ | 412.1084 |
| 1987 | 165.8 | $X_0$ | 187.6464 | 2007 | 430.8 | $X_2$ | 412.1084 |
| 1988 | 179.7 | $X_0$ | 187.6464 | 2008 | 457.1 | $X_2$ | 412.1084 |
| 1989 | 194.6 | $X_0$ | 187.6464 | 2009 | 492.8 | $X_2$ | 412.1084 |
| 1990 | 200.9 | $X_0$ | 187.6464 | 2010 | 532.0 | $X_3$ | 540.8059 |
| 1991 | 211.1 | $X_0$ | 187.6464 | 2011 | 532.7 | $X_3$ | 540.8059 |
| 1992 | 229.3 | $X_1$ | 257.7034 | 2012 | 540.0 | $X_3$ | 540.8059 |
| 1993 | 238.3 | $X_1$ | 257.7034 | 2013 | 556.4 | $X_3$ | 540.8059 |
| 1994 | 246.0 | $X_1$ | 257.7034 | 2014 | 565.8 | $X_3$ | 540.8059 |
| 1995 | 253.8 | $X_1$ | 257.7034 | 2015 | 609.2 | $X_3$ | 540.8059 |
| 1996 | 270.1 | $X_1$ | 257.7034 | 2016 | 639.2 | $X_3$ | 540.8059 |
| 1997 | 285.7 | $X_1$ | 257.7034 | 2017 | 657.8 | $X_3$ | 540.8059 |
| 1998 | 295.9 | $X_1$ | 257.7034 | 2018 | 675.4 | $X_3$ | 540.8059 |
| 1999 | 292.3 | $X_1$ | 257.7034 | 2019 | 728.7 | $X_3$ | 540.8059 |
| 2000 | 300.0 | $X_1$ | 257.7034 | 2020 | 730.8 | $X_3$ | 540.8059 |

### B. Forecasting method II: Forecasting model based on Chen's arithmetic model

The production forecasting of the Coal is based on the 40 years (1980-2020).

**Table 3.** Coal production forecast

| Year | Actual Production | Forecasted method | Year | Actual Production | Forecasted method |
|---|---|---|---|---|---|
| 1981 | 113.9 | 200 | 2001 | 309.6 | 400 |
| 1982 | 124.2 | 200 | 2002 | 327.8 | 400 |
| 1983 | 130.5 | 200 | 2003 | 341.3 | 400 |
| 1984 | 138.2 | 200 | 2004 | 361.3 | 400 |
| 1985 | 147.4 | 200 | 2005 | 382.6 | 400 |
| 1986 | 154.2 | 200 | 2006 | 407.0 | 500 |
| 1987 | 165.8 | 200 | 2007 | 430.8 | 500 |
| 1988 | 179.7 | 200 | 2008 | 457.1 | 500 |
| 1989 | 194.6 | 200 | 2009 | 492.8 | 500 |
| 1990 | 200.9 | 300 | 2010 | 532.0 | 600 |
| 1991 | 211.1 | 300 | 2011 | 532.7 | 600 |

| | | | | | |
|---|---|---|---|---|---|
| 1992 | 229.3 | 300 | 2012 | 540.0 | 600 |
| 1993 | 238.3 | 300 | 2013 | 556.4 | 600 |
| 1994 | 246.0 | 300 | 2014 | 565.8 | 600 |
| 1995 | 253.8 | 300 | 2015 | 609.2 | 700 |
| 1996 | 270.1 | 300 | 2016 | 639.2 | 700 |
| 1997 | 285.7 | 300 | 2017 | 657.8 | 700 |
| 1998 | 295.9 | 300 | 2018 | 675.4 | 700 |
| 1999 | 292.3 | 300 | 2019 | 728.7 | 750 |
| 2000 | 300.0 | 300 | 2020 | 730.8 | 750 |

**Table 4.** A Comparison of MSE

| | Forecasted Method-1 | Forecasted Method-2 |
|---|---|---|
| MSE | 19,015.34783 | 60,000.516 |

## 4. CONCLUSION

This paper presents the two methods for fuzzy time series forecasting. The method has been implemented on the historical time series data of coal production of India to have a comparative study with the existing methods. The mean square error of the k-means clustering method is 19,015.34783 and the mean square error of the chen's arithmetic method is 60,000.516. From Table 4 we can see that the forecasted method 1 has a higher forecasting accuracy rate than the forecasted method 2. After examining, the k-means clustering method was the best and most appropriate to apply to study data series of coal production.

## REFERENCES

[1] Chen S. M, (1996), "Forecasting enrollments based on fuzzy time series", Fuzzy Sets and Systems, 81, pp:311-319.

[2] DuruOken, Bulut Emrahand Yoshida Shigeru(2010): "Bivariate Long Term Fuzzy Time Series Forecasting of DryCargo Freight Rates", The Asian Journal of shipping andLogistics,Vol.28,No.2,pp.207-223.

[3] Huarng.K, (2001), "Heuristic models of fuzzy time series for forecasting", Fuzzy Sets and Systems, 123, pp:369-386

[4] Jilani. T.A., et.al., (2009),"Fuzzy metric approach for fuzzy time series forecasting based on frequency density-based partitioning", In: Proceedings of World Academy of Science, Engineering and Technology 23,pp:1307-6884.

[5] Ravi, K., Vadlamani, R., Prasad, P., "Fuzzy formal concept analysis-based opinion mining for CRM in financial services". Appl. Soft Comput. 58, (2017), 35–52.

[6] Song.Q, Chissom . B.S, (1993b), "Forecasting enrollments with fuzzy time series—Part I", Fuzzy Sets and Systems, 54, pp:1-10.

[7] Song.Q, Chissom .B.S, (1994) ,"Forecasting enrollments with fuzzy time series—Part II", Fuzzy Sets and Systems, 62 , pp:1-8.

[8] Wang.J. R. H, et.al., (1998) ":Handing forecasting problems using fuzzy time series", Fuzzy Sets and Systems, 100 ,217-228.

[9] Yu,Hui-Kuang, "A refined fuzzy time series model for forecasting". Physica A. 346[2005], 657-681.

[10] Zhigiangzhang (2012),"Fuzzy time series forecasting based on k-means Clustering",open journal applied sciences, pp:100-103.