INTERNATIONAL JOURNAL FOR INNOVATIVE RESEARCH IN MULTIDISCIPLINARY FIELD
ISSN(O): 2455-0620 [ Impact Factor: 9.47 ]
Monthly, Peer-Reviewed, Refereed, Indexed Journal with IC Value : 86.87
Volume - 12, Issue - 3, March - 2026

DOIs:10.2015/IJIRMF/202603015 --:-- Research Paper / Article / Review

# A Comparative Study of Closest Fit Approaches to Recover Missing Attributes Values

**Dr Sanjay Gour**
Professor, Gandhinagar University, Gandhinagar, Gujrat
Email – sanjay.since@gmail.com

***Abstract:*** *The Journey of data and its mining with analysis is endless, the basic types of data are numerical and character, other are hybrid. Usually, dataset or attributes with missing values confuses both the data mining and data analysis. It also affects the application of the result to novel data as well as concluding result. It also generated ambiguities in the datasets, need to treat before applying data mining, analysis or machine learning model. In order to recover missing values, earlier closest fit approaches are utilized during the data cleaning phase of pre-processing, resultant in excellent result with linear-numeric dataset. The present study is a comparison between three key closest fit approaches and their consequences. It is significant review and comparative study of these approaches and their all-time applicability in the domain of data mining, analysis and machine learning.*

***Key Words:*** *Data Mining, Missing Values, Attribute, Data preparation, Closest fit.*
MSC (2010) Subject Classification: 62-07,62N02, 62Q99

## 1. INTRODUCTION:

The proposed comparison study of closest fit approaches is basically a review of consequences received from the three closest fit approaches namely simple average closest fit, closest fit and segmented closest approach to recover the missing attributes values, utilize in the data mining and analysis as data pre-processing. These methods are using artificially generated values to replace the missing values. These methods are suitable for numerical attributes and is exploration of closest fit value, very near to the mean of the attribute and closest to impartial preceding and succeeding values.

The function of statistical methods has gained stuff in exploring estimation and prediction techniques. Allison [1] explored approximations about linear models with missing values and incomplete data. Buck [2] recommended approximation of missing values for utilization with an electronic computer. Chen et. Al [3] considered and deliberated about several imputation for missing data. Darshanaben et al. [4] discussed the method to treat the anomalies values which will be further converted into missing values before handling. Gaur and Dulawat [5,6,7] work on a series of algorithms based on closest fit approach which are valuable for approximation of missing values, provides closest fit approach with segmented, univariate and improved approach analysis by using central tendency at the location of missing values for data cleaning. Gyzymala-Busse [8] spring clue that each missing dataset values is substituted by entire probably identified values, also worked on the concept of closest fit technique to treat the missing values. Rubin [9] discovered the statistical inference to deal the missing values with multiple imputations in order to response the non-responses during survey. S. Gaur et al. [10, 11] discussed the statistical inference and missing data handling with data mining. Swati et al. [12] explore the concept of agile method to manage missing block in data mining. Sanjay et al. [13] discussed the applied interpolation method for recover randomly missing values in data mining.

## 2. SIMPLE AVERAGE CLOSEST FIT APPROACH:

This is initial approach and utilised as the usual substitute to manage missing values in numerical attributes. It is based on the substituting missing values via average generated value, useful for flag of linear analysis. Usually, simple average closest fit approach is federal on search of value, very close towards attributes central tendency and aligned with succeeding and preceding value of the missing values. Followings are the steps:

```
Read     X = {x_1, ..., x_n}   // Attribute with observed and missing values
         where X = X_obs + X_mis
         X_obs = {x_1, ..., x_k}        // Attribute values observed
         X_mis = {x_{k+1}, ..., x_n}    // Attribute values missing
For i = 1 to n do
    If ( value (x_i) == NULL) then
        x_p = value(x_{i-1})    // Value of preceding of x_i
        x_s = value(x_{i+1})    // Value of succeeding of x_i
        x̄_ps = (x_p + x_s) / 2  // Average of preceding and succeeding
        x_est = x̄_ps            // Estimated value
        value (x_i) = x_est     // Assigning estimated value to missing
                                   value place
    i = i + 1
    repeat until (i >=n)
    Stop
```

Figure 1: Algorithm for simple average closest fit approach

## 3. CLOSEST FIT APPROACH

This is the first method in the series of closest fit approach, very much valuable for numerical attributes. As a whole, this process is quest of finding closest fit value. It means about true average of the attribute in association of nearest or neighbouring to the succeeding and preceding value of the missing values. To generate the closest fit values there is three step process adopted are as follows:

1. Find out the mean of whole attribute on the current observed values denoted as:
$$\bar{X}_{obs}=\frac{1}{k}\sum_{i=1}^{k} x_i$$
2. Calculate the central tendency of preceding and successing values of the missing value which is denoted as:
$$\bar{x}_{ps} = (x_p + x_s)/2$$
3. Find the values from process 1 and 2. Calculate the mean values from them which is denoted as the Estimated value.
$$x_{est} = (\bar{X}_{obs} + \bar{x}_{ps}) / 2$$
4. Post the estimated value at the location of missing value.
$$\text{value } (x_i) = x_{est}$$

## 4. SEGMENTED CLOSEST FIT APPROACH

It is advance step of closest fit approach or we can say the multi-fold implementation of the same. In this approach segmented closest fit value is generated as replacement of missing value, first measure the length of the attribute then compute the $1/5^{th}$ (20%) of the attribute-length. The 20% parts of the attribute is acknowledged as size of slices or attribute segments; thus, attribute is segmented into five equal shares. Respectively missing value will be part of any one of the segments. To estimate the missing value following procedure will be followed:

1. First create the segments of the given attribute.
2. Find the segment from where the missing value gets located or observed.
3. Calculate the arithmetic mean of segment in which missing values location is noted. It is denoted as (for first segment):
$$\bar{x}_{obs} = \frac{1}{m}\sum_{i=1}^{m} x_i$$
For the second segment it will be as:
$$\bar{x}_{obs}=\frac{1}{m}\sum_{i=m+1}^{2m} x_i$$
4. Calculate the central tendency of preceding and successing values of the missing value which is denoted as:

$$\bar{x}_{ps} = (x_p + x_s)/2$$

5. Find the values from process 3 and 4. Calculate the mean values from them which is denoted as the Estimated value.

$$x_{est} = (\bar{X}_{obs} + \bar{x}_{ps}) / 2$$

6. Apply step 3, 4 and 5 separately for each segment.
7. Post the estimated value at the location of missing value.

$$\text{value } (x_i) = x_{est}$$

## 5. THE DATA SOURCE

The datasets for the study were occupied from the web site of earth policy "www.earth_policy.org", purely academic purpose. The time series-cross sectional dataset was considered for study, their consistent years are from 1960 to 2009, consumption of Coal, Oil and Natural gas globally. The size of datasets considered for the study was as, total number of values in each attribute (Coal, Oil and Natural Gas) are 50 denoted as standard dataset. The variant of such datasets was as with missing values case and Recovered values case. So, a dataset passes thru three phases of manipulations.
**Dataset for Experimental**: The missing values case were considered in two variants as 05% and 10% missing values dataset. So, total type of datasets was:

- Standard dataset with complete values in the attribute
- Dataset with 05% randomly missing values.
- Dataset with 10% randomly missing values
- Dataset with 05% recovered values as replacements of missing values.
- Dataset with 10% recovered values as replacements of missing values.

## 6. METHODOLOGY

The observation for the study was taken according to the experimental datasets discussed above. The process was as:

- Read the complete dataset and keep record with size, mean, standard deviation (SD) and coefficient of variation (CV).
- Missing 05% and 10% values in planned way with random missing approach, calculate mean, standard deviation (SD) and coefficient of variation (CV).
- Recover the missing 05% and 10% values with estimated values, calculate mean, standard deviation (SD) and coefficient of variation (CV).
- Perform the analysis process.

## 7. OBSERVATION

Followings are observations from the experimental with diverse missing values cases and recoveries.

**Simple Average Closest Fit Approach:**

| Variable | Standard | 05% missing | 10% missing | 05% Recovered | 10% Recovered |
|---|---|---|---|---|---|
| Coal | 2109 | 2113 | 2111 | 2109 | 2109 |
| Oil | 2262 | 2248 | 2250 | 2261 | 2264 |
| Gas | 879 | 884 | 874 | 879 | 879 |

Table 1 (a): Observations from simple average closest fit approach

**Closest Fit Approach:**

| Variable | Standard | 05% missing | 10% missing | 05% Recovered | 10% Recovered |
|---|---|---|---|---|---|
| Coal | 2109 | 2113 | 2111 | 2111 | 2110 |
| Oil | 2262 | 2248 | 2250 | 2255 | 2257 |
| Gas | 879 | 884 | 874 | 881 | 877 |

Table 1 (b): Observations from Closest fit approach

**Segmented Closest Fit Approach:**

| Variable | Standard | 05% missing | 10% missing | 05% Recovered | 10% Recovered |
|---|---|---|---|---|---|
| Coal | 2109 | 2113 | 2111 | 2107 | 2110 |
| Oil | 2262 | 2248 | 2250 | 2255 | 2257 |
| Gas | 879 | 884 | 874 | 881 | 879 |

Table 1 (c): Observations from Segmented closest fit approach

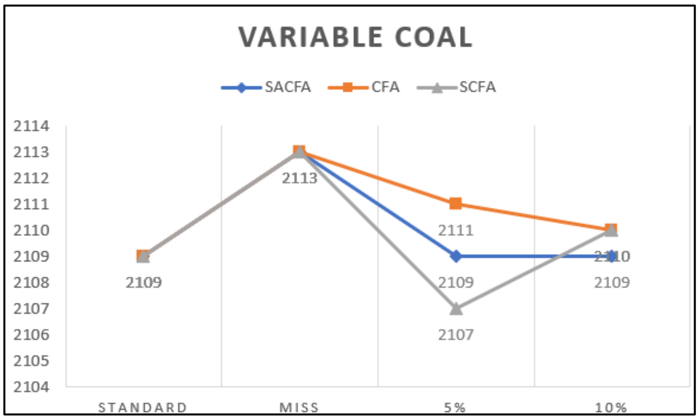

Figure 2(a): Observations from variable Coal

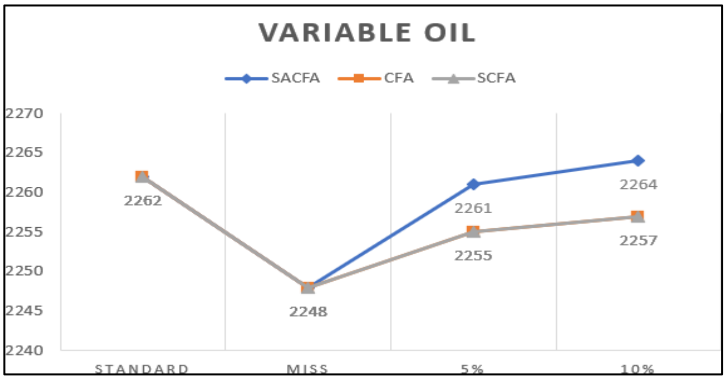

Figure 2(b): Observations from variable Oil

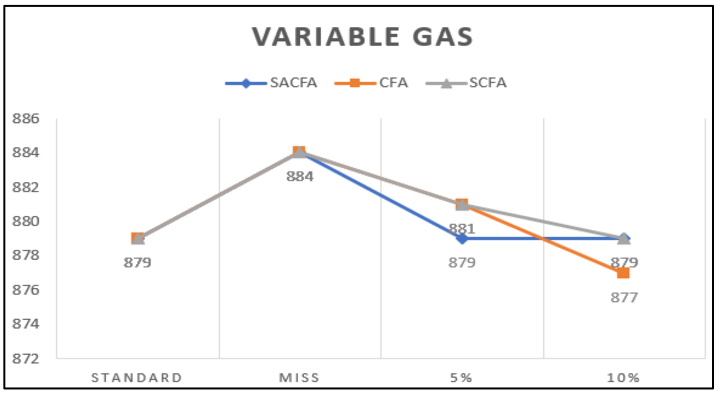

Figure 2(c): Observations from variable Natural Gas

From the table 1(a), (b) and (c) and Figure 2 (a), (b)and (c) it is observed that:

- In the variable coal it is observed that almost result is similar, SACFA gives 100% recovery at both levels. CFA and SCFA both are slightly lower than SACFA, which are almost negligible as deviation are 0.095 and 0.047.
- In the variable Oil also observed that almost result is similar, at 10% recovery level CFA and SCFA are performing slightly better than SACFA.
- In the natural gas variable also observed that almost result is similar, here at 10% level SACFA and SCFA recovered 100% where CFA has minor deviation which is negligible.

**8. DEVIATION AND PERFORMANCE:**

To justify the performance in minute manner we are classifying performance on the basis of result of recovery result of 5% and 10% missing values. The performance grade given with a 3-scale numeric values as 3) Up, 2) Mid and 1) Below. Thus, the matrix of performance after scale-based calculations the conclusions are as given below:

| Coal | Standard | 5% | Deviation | 10% | Deviation | Performance |
|---|---|---|---|---|---|---|
| SACFA | 2109 | 2109 | 0.000 | 2109 | 0.000 | Up |
| CFA | 2109 | 2111 | 0.095 | 2110 | 0.047 | Mid |
| SCFA | 2109 | 2107 | -0.095 | 2110 | 0.047 | Mid |

Table 2 (a): Deviation and performance table with variable Coal

| Oil | Standard | 5% | Deviation | 10% | Deviation | Performance |
|---|---|---|---|---|---|---|
| SACFA | 2262 | 2261 | -0.044 | 2264 | 0.088 | Below |
| CFA | 2262 | 2255 | -0.309 | 2257 | -0.221 | Mid |
| SCFA | 2262 | 2255 | -0.309 | 2257 | -0.221 | Mid |

Table 2 (b): Deviation and performance table with variable Oil

| Gas | Standard | 5% | Deviation | 10% | Deviation | Performance |
|---|---|---|---|---|---|---|
| SACFA | 879 | 879 | 0.000 | 879 | 0.000 | Up |
| CFA | 879 | 881 | 0.228 | 877 | -0.228 | Below |
| SCFA | 879 | 881 | 0.228 | 879 | 0.000 | Mid |

Table 2 (c): Deviation and performance table with variable Gas

| Approach | Coal | Oil | Gas | Total |
|---|---|---|---|---|
| **SACFA** | 3 | 1 | 3 | 7 |
| **CFA** | 2 | 2 | 1 | 5 |
| **SCFA** | 2 | 2 | 2 | 6 |

Table 3: Performance table

## 9. RESULT AND CONCLUSIONS

It is observed that there is no major difference regarding the performance of approaches from the numerical calculation, it is found that result is very close to each other. Although simple average closest fit approach showing highest performance, Segmented closest fit approach slightly below to SACFA and closest fit approach is slightly below to Segmented closest fit approach. From the size of the dataset, it is observed that SACFA is very good approach for time series and one of the key approaches, but it gives good result only for time series attribute. The closest fit gives best result for balances and linear kind of attributes. The segmented closest fit approach is sometimes better than all other when distribution of data are not equal in the attribute. It also best to high volume attributes.

## REFERENCES:


1. Allison, P.D., (2001). Missing data, Thousand Oaks CA: *Sage publication*, (2001).
2. Buck, S.F., (1960): A method of estimation of missing values in multivariate data suitable for use with an electronic computer, *J. Royal Statistical Society, Series* B, 2, 302-306.
3. Chen, L., Drane, M.T., Valois, R.F., and Drane, J.W., (2005): Multiple imputation for missing ordinal data, *Journal of Modern Applied Statistical Methods, 4(1), 288-299.*
4. Darshanaben Dipakkumar Pandya and Sanjay Gaur, (2018): Detection of Anomalous value in Data Mining, *Kalpa Publications in Engineering, 2, 1-6.*
5. Gaur, Sanjay and Dulawat, M.S., (2011): Improved closest fit techniques to handle missing attributes values, *Journal of Computer and Mathematical Sciences,* 2(2), 384-390.
6. G Sanjay, M S Dulawat,(2011): Segmented Closest Fit Approach to Handle Missing Data, *Ultra Scientist of Physical Sciences,* 23(2).
7. Gaur, Sanjay and Dulawat, M.S. (2010): Univariate analysis for data preparation in context of missing values, *Journal of Computer and Mathematical Sciences*, 1(5), 628-635.
8. Grzymala-Busse, J. W., (2004): Data with missing attribute values: Generalization of in-discernibility relation and rules induction, Transactions of Rough Sets, *Lecture Notesin Computer Science Journal Subline, Springer-Verlag*, 1, 78-95.
9. Rubin, D.B., (1976): Inference and missing data, *Biometrika*, 63, 581-592.
10. S. Gaur and M.S. Dulawat (2010): A perception of statistical inference in data mining, *International Journal of Computer Science and Communication,* 1(2), 653-658.

11. S Gaur, DD Pandya, D Soni, (2019), Closest fit approach through linear interpolation to recover missing values in data mining, Fourth International Congress on Information and Communication Technology: ICICT 2019, London, 1, 513-521. Springer Singapore
12. S Swati, G Sanjay, (2013): Contiguous Agile Approach to Manage Odd Size Missing Block in Data Mining, *International Journal of Advanced Research in Computer Science* 4 (11), 214-217.
13. Sanjay G., Darshanaben D. P., Manish K. S., (2019) Applied NF interpolation method for recover randomly missing values in data mining, Fourth International Congress on Information and Communication Technology: ICICT 2019, London, 2, 475-485, Springer Singapore.