SPU Journal of Science, Technology and Management Research (SPU-JSTMR) Special Issue-ICAICS-26
ISSN:3049-1479(Online)

# An Analytical Framework for Reconstructing Missing Time-Series Data Using Interpretable AI Methods

**Sneh B. Patel[1] and Dr. Darshanaben D. Pandya[2]**

Research Scholar, Department of Computer Application, Sankalchand Patel University, Visnagar, Gujarat, India[1]
Associate Professor, Department of Computer Science & Application, Shri C. J Patel College of Computer Studies (BCA), Sankalchand Patel University, Visnagar, Gujarat, India[2]

spdigitalme@gmail.com,ddpandya.fcs@spu.ac.in

**Abstract:** Large AI systems and foundation models rely on continuous and consistent numeric data. In real life, missing values are common in the time-series data because of gaps in reporting, measurement failures, or archival errors. If these are not handled well, they have the potential to alter the distribution of the data, bias learning, and hurt model performance. In this paper, we presented a simple, intuitive, deterministic interpolation approach to fill gaps in long-term numeric time series. Unlike the imputation approach, which has latent, black-box-type imputation, this uses a step-by-step mathematical process; as such, tracing and verifying statistically are easily achieved. We tested the method on a historical Standard Yield (SY) dataset from 1950–2009, to which we artificially introduce missing values and then reconstruct. We utilize a comprehensive set of statistical tests for the descriptive statistics, variance and covariance comparisons, and one-way ANOVA to demonstrate that in average value, the dispersion, and the time-varying process, the original data is closely matched with the recovered series. This demonstrates that classic, interpretable reconstruction methods are still relevant in a modern AI pipeline, during the preprocessing stage of foundation models, which demands data quality, stability, and explain ability.

**Keywords:** Time-series reconstruction, missing value estimation, interpretable AI, foundation models, statistical validation

## I. INTRODUCTION

Foundation models constitute the current central concept in artificial intelligence [2]. They are improved by having more data, and they can make use of what they learn across diverse applications. Their success depends on the quality and coherence of the training data [3]. Even small incompleteness and poorly filled values can introduce bias, instability during training, and reduce generalization performance of the model.

Missing values are common in time-series data applied to environment, economy, and policy work, as data is usually collected over a long period and from different sources [10]. New AI-based gap-filling methods can be flexible but often incomprehensible, demand higher computational resources, and cannot be checked for flaws easily. On the other hand, methods based on transparent mathematics are reproducible, and their behavior is predictable.

It is an exploration of how such a cautious, interpolation-based approach can fill in missing values while preserving the big-picture statistics required to make robust AI models [1]. This work has two complementary aspects: (i) the actual filling of missing data in a real historical dataset, and (ii) the implications of this filling for foundation model data preparation.

## II. LITERATURE REVIEW

### 1.1 Missing Data in Time-Series and AI Applications

There are different patterns of missing data: MCAR, MAR, and MNAR; these affect the quality with which gaps are filled. Missing values in time series are rarely random and could affect how generalizable AI models learn from the data [4]. Recent studies have shown that anomalous number sequences could hurt embedding stability and speed up large model training.


**http://doi.org/10.63766/spujstmr.26.000069**

### 2.2 Classical Deterministic Imputation Methods

Old-school approaches- such as imputation by mean, regression imputation, and interpolation-are widely used because they are easy to interpret. Linear and polynomial interpolationassume smooth changes over time and work when trends are stable. They can miss sharp changes, but as they are deterministic, they are much easier to explain and audit in AI processes.

### 2.3 AI-Based and Hybrid Imputation Techniques

Machine-learning-based imputers, such as k-NN, random forests, autoencoders, and diffusion models, can catch complex links but often require lots of data, tuning, and computing power. They can also often act like a black box, making interpretation difficult-a factor many AI rules and checks wish to avoid [8].

### 2.4 Research Gap

Even though there are numerous studies on advanced imputers, few pieces of work have regarded how deterministic interpolation would affect the statistical quality of data used in foundation models [9]. Few papers verify how well the variance is preserved, the structure of covariance, or how consistent the hypothesis tests are after gap filling. This paper looks to fill this gap.

## III. METHODOLOGY

### 3.1 Dataset Acquisition

The underlying time-series of Standard Yield (SY) covers the period from 1950 to 2009. The standard dataset, SY_Standard, contains 60 annual observations.

TABLE I
SY DATASET WITH STANDARD, MISSING AND RECOVERED VALUES

| Year | SY_Standard | SY_Missing | SY_Recover |
|---|---|---|---|
| 1950 | 1.46 | 1.46 | 1.46 |
| 1951 | 1.40 | 1.40 | 1.40 |
| 1952 | 1.39 | 1.39 | 1.39 |
| 1953 | 1.22 | 1.22 | 1.22 |
| 1954 | 1.35 | 1.35 | 1.35 |
| **1955** | 1.35 | | **1.54** |
| 1956 | 1.47 | 1.47 | 1.47 |
| 1957 | 1.56 | 1.56 | 1.56 |
| 1958 | 1.63 | 1.63 | 1.63 |
| 1959 | 1.58 | 1.58 | 1.58 |
| 1960 | 1.58 | 1.58 | 1.58 |
| 1961 | 1.69 | 1.69 | 1.69 |
| 1962 | 1.63 | 1.63 | 1.63 |
| 1963 | 1.64 | 1.64 | 1.64 |
| 1964 | 1.53 | 1.53 | 1.53 |
| 1965 | 1.65 | 1.65 | 1.65 |
| 1966 | 1.71 | 1.71 | 1.71 |
| 1967 | 1.65 | 1.65 | 1.65 |
| **1968** | 1.80 | | **1.74** |
| 1969 | 1.84 | 1.84 | 1.84 |
| 1970 | 1.79 | 1.79 | 1.79 |
| 1971 | 1.85 | 1.85 | 1.85 |
| 1972 | 1.87 | 1.87 | 1.87 |

| | | | |
|---|---|---|---|
| 1973 | 1.87 | 1.87 | 1.87 |
| 1974 | 1.59 | 1.59 | 1.59 |
| 1975 | 1.94 | 1.94 | 1.94 |
| 1976 | 1.75 | 1.75 | 1.75 |
| 1977 | 2.06 | 2.06 | 2.06 |
| 1978 | 1.97 | 1.97 | 1.97 |
| 1979 | 2.16 | 2.16 | 2.16 |
| 1980 | 1.78 | 1.78 | 1.78 |
| **1981** | 2.02 | | **2.02** |
| 1982 | 2.12 | 2.12 | 2.12 |
| 1983 | 1.76 | 1.76 | 1.76 |
| 1984 | 1.89 | 1.89 | 1.89 |
| 1985 | 2.29 | 2.29 | 2.29 |
| 1986 | 2.24 | 2.24 | 2.24 |
| 1987 | 2.28 | 2.28 | 2.28 |
| 1988 | 1.82 | 1.82 | 1.82 |
| 1989 | 2.17 | 2.17 | 2.17 |
| 1990 | 2.29 | 2.29 | 2.29 |
| 1991 | 2.30 | 2.30 | 2.30 |
| 1992 | 2.53 | 2.53 | 2.53 |
| 1993 | 2.19 | 2.19 | 2.19 |
| **1994** | 2.78 | | **2.31** |
| 1995 | 2.38 | 2.38 | 2.38 |
| 1996 | 2.53 | 2.53 | 2.53 |
| 1997 | 2.62 | 2.62 | 2.62 |
| 1998 | 2.62 | 2.62 | 2.62 |
| 1999 | 2.46 | 2.46 | 2.46 |
| 2000 | 2.56 | 2.56 | 2.56 |
| 2001 | 2.66 | 2.66 | 2.66 |
| 2002 | 2.56 | 2.56 | 2.56 |
| **2003** | 2.28 | | **2.67** |
| 2004 | 2.84 | 2.84 | 2.84 |
| 2005 | 2.90 | 2.90 | 2.90 |
| 2006 | 2.88 | 2.88 | 2.88 |
| 2007 | 2.81 | 2.81 | 2.81 |
| 2008 | 2.67 | 2.67 | 2.67 |
| 2009 | 2.91 | 2.91 | 2.91 |

### 3.2 Missing-Value Simulation

We created a controlled environment to evaluate how well we can fill in missing data by removing values from years that show stable trends and do not change much. The years whose data have been removed include:

- 1955, 1968, 1981, 1994, and 2003

This leaves SY_Missing with fifty-five valid observations and five missing values. The positions of the missing values were selected to reflect realistic data gaps often encountered in environmental and socio-economic historical time-series data.

### 3.3 Interpolation-Based Recovery Algorithm

To fill in the missing values, we built a custom interpolation method. It estimates the missing value $y_k$at point$x_k$ using a few intermediate calculations. It is based on linear interpolation but given as a clear step-by-step procedure.

Given known pairs:

- $x = [x_1, x_2, \dots, x_n]$,
- $y = [y_1, y_2, \dots, y_n]$,

and a missing point $x_k$, do the following:

**Step-by-Step Computational Procedure**
Step 1: Compute the product list $xy$

$xy_i = x_i * y_i$
$xy_i = xi \cdot yi$

Step 2: Determine range components
$x\_min = \min(x)$
$x\_max = \max(x)$
$\Delta x = x\_max - x\_min$

Step 3: Compute sum of xy products
$S = \Sigma (x_i * y_i)$

Step 4: Compute $v_1$
$v_1 = 1 + \Delta x$

Step 5: Compute $v_2$
$v_2 = S - (\Delta x * x_k)$

Step 6: Compute $v_3$
$v_3 = v_2 / \Delta x$

Step 7: Compute $v_4$
$v_4 = v_1 * v_3$

Step 8: Compute $v_5$
$v_5 = v_4 - v_2$

Step 9: Compute $v_6$
$v_6 = v_5 / x_k$

Step 10: Final reconstructed value
$v_7 = v_6 + 1$

The final recovered SY for year $x_k$ is $v_7$.
This step-by-step method helps keep the calculation clear and easy to interpret, which is important when adding it to data-preprocessing workflows for foundation models.

### 3.4 Statistical Evaluation

We examined the reconstructed series using:

- Descriptive statistics (mean, variance, standard deviation)
- One-way ANOVA
- Covariance matrix analysis

# IV. RESULTS

This section presents the result of reconstructing the SY dataset, by comparing three data series: the SY_Standard, SY_Missing, and SY_Recover datasets, with regard to the stability of the distribution pattern of data, retention of variation, and structure similarity, using descriptive statistics, analysis of variance, and covariance.

### 4.1 Descriptive Statistics

TABLE II

DESCRIPTIVE STATISTICAL COMPARISON OF SY_STANDARD, SY_MISSING, AND SY_RECOVER DATASETS

| Statistic | SY_Standard | SY_Missing | SY_Recover |
|---|---|---|---|
| Mean | 2.04 | 2.03 | 2.04 |
| Variance | 0.22 | 0.22 | 0.21 |
| Count | 60 | 55 | 60 |

The recovered series is, in fact, very close to the original distribution.

### 4.2 One-Way ANOVA

A one-way ANOVA has been conducted to compare the rebuilt series against the original one using three datasets: SY_Standard, SY_Missing, and SY_Recover. The null hypothesis states that all series have the same population mean.

TABLE III

ONE-WAY ANOVA RESULTS FOR SY_STANDARD, SY_MISSING, AND SY_RECOVER DATASETS

| Source of Variation | Sum of Squares (SS) | Degrees of Freedom (df) | Mean Square (MS) | F-value | p-value |
|---|---|---|---|---|---|
| Between Groups | 0.00018 | 2 | 0.00009 | 0.00015 | 0.999985 |
| Within Groups | 6.06042 | 177 | 0.03425 | — | — |
| Total | 6.06060 | 179 | — | — | — |

### 4.3 Covariance Matrix Analysis

While ANOVA looks at averages, covariance analysis shows how well the trends fit, and how similar the structures are. High covariance between the recovered and original series means the time pattern is well kept.

TABLE IV
COVARIANCE MATRIX OF SY_STANDARD, SY_MISSING, AND SY_RECOVER DATASETS

| Series Pair | Covariance |
|---|---|
| SY_Standard ↔ SY_Standard | 0.2144 |
| SY_Standard ↔ SY_Missing | 0.2141 |
| SY_Standard ↔ SY_Recover | 0.2136 |
| SY_Missing ↔ SY_Missing | 0.2168 |
| SY_Missing ↔ SY_Recover | 0.2172 |
| SY_Recover ↔ SY_Recover | 0.2116 |

### 4.4 Reconstruction Accuracy and Pattern Preservation

Beyond numerical results and formal tests, the reconstruction values should be evaluated for how well they integrate into the temporal history of the dataset. In the context of this, reconstruction accuracy will be defined not as an exact point-wise prediction but rather as the preservation of structural coherence, smoothness, and temporal realism.

The stated criteria would recover in the following years: 1955, 1968, 1981, 1994, and 2003.

- **Trend consistency:** every reconstructed value stays within the neighborhood range without large jumps or flat spots.
- **Temporal Continuity:**interpolationkeeps a smooth year-to-year trend and fills all time points without gaps.
- **Distributional Alignment:**Recovered points do not artificially change the overall variability - this is reflected in the descriptive statistics and covariance analyses.
- **No Over-Smoothing:**this methodology does not over-smooth; it retains all the ups and downs in their natural shape with different spline and high-order fitting methods.

From a foundation-model view, the most important points are keeping the temporal pattern helps realistic dynamics, which is crucial for embedding stability, attention alignment, and temporal generalization in time-aware AI. Generally, this deterministic interpolation can achieve high structural accuracy and will be useful for historical data restoration used in the training, test, and evaluation of AI.

## V. DISCUSSION

These experiments demonstrate that this interpolation-based reconstruction method effectively fills in missing values in the SY time-series data without hurting its overall statistical quality [6]. This is especially important for AI systems and foundation models, which require complete, stable, and structurally consistent data in the pre-training and evaluation. The results also showed that a transparent, deterministic reconstruction can meet the current AI data-quality requirements without adding bias or instability.

### 5.1 Preservation of Statistical Characteristics

There is a strong agreement between the original series (SY_Standard) and the recovered series (SY_Recover). They share the same mean of 2.04 and very similar variances; thus, reconstruction does not affect the data distribution, nor does it introduce bias [5]. The missing data, SY_Missing, is characterized by a lower variance, which can show how gaps without filling can mask the true variability. Such variability is fully recovered by the recovered series, proving that the method keeps realistic data behavior.

### 5.2 ANOVA-Based Structural Validation

The results of one-way ANOVA support that these datasets are statistically the same. The very high p-value 0.999985 shows no real difference among SY Standard, SY_Missing, and SY_Recover. Recovered values are statistically indistinguishable from the original data in terms of central tendency, which, for foundation models, matters to prevent drift or bias during embedding learning or optimization.

### 5.3 Covariance and Trend Fidelity

Covariance analysis shows that the recovered series retains the time pattern of the original data. The covariance of SY Standard and SY_Recover is very close to the self-covariance of SY Standard; therefore, the trends are highly aligned with the patterns. This, for its part, ensures that the recovered values will fit naturally into the historical path of the dataset, something quite relevant in time-series modeling and AI-based representation learning.

### 5.4 Implications for Foundation Model Development

These results are important for creating data to be used in training foundation models. The reconstruction method, by keeping the average and spread and the time patterns, helps the data quality and stability during training, and reduces the possibility of bias and shifts in the data due to missing or poorly filled values. Unlike deep learning fillers, this approach is intuitive, easy to understand, and quick in terms of runtime because each step is clearly described and therefore simple to follow. This improvesreproducibility, auditability, and responsible use of AI [7]. Its light design fits large data preparation jobs for environmental, economic, and policy time series data when consistency and speed are important.

## VI. FUTURE WORK

Might combine deterministic interpolation with either probabilistic modeling or learning-based temporal representations. Such extensions may achieve greater flexibility and accuracy with no loss in interpretability and statistical stability shown here.

### 5.6 Summary of Findings

On the whole, the method reliably fills in gaps in values while preserving the essential statistical and temporal properties of the original data. Please feel free to contact us today with questions or if you would like to learn more about our firm.

## VII. CONCLUSION

This work presented an intelligible method to impute missing values in numerical, long-term time series data with interpolation, suitable for modern AI and foundation model workflows. Using a historical dataset of the SY from 1950 to 2009, this paper correctly imputed missing observations while preserving relevant statistical properties such as mean, variance, covariance, and overall pattern in time. Statistical checks confirmed that the reconstructed series is of a similar appearance as the original data. No major differences were present in the ANOVA tests; covariance matched well, suggesting that the approach keeps the distribution and trend of data. This supports that the method decreases the problems induced by missing data without creating bias or altering the structure of data.

In practice, the approach presents a straightforward and lightweight counterpart to the fancy learning-based imputers. Because of its deterministic nature, it is easy to understand, reproduce, and fit into extensive pre-processing steps. As foundation models rely more on good numerical and time-series data, these qualities help make sure that training is stable and results are dependable. Taken together, these results show the value of transparent, classical reconstruction techniques in modern AI systems. This research thus contributes to the development of robust, interpretable, and data-efficient foundation models in science and engineering by showing that missing values can be recovered accurately and reliably using only simple math.

### References

[1] OECD. (2021). *Artificial intelligence, big data and the future of skills*. OECD Publishing. https://doi.org/10.1787/9789264314626-en

[2] Stanford University. (2022). *Artificial Intelligence Index Report 2022*. Human-Centered AI Institute, Stanford University.

[3] Brynjolfsson, E., & McAfee, A. (2017). *The second machine age: Work, progress, and prosperity in a time of brilliant technologies*. W. W. Norton & Company.

[4] Pandya, D.D., Jadeja, A., Trivedi, S., Patel, P.A., Tamhankar, I., Dhanesha, P.S. (2026). Internet of Things (IoT)'s Transformative Power Enhancing Wireless Networks and Sensing. In: Rathore, V.S., Piuri, V., Babo, R., Karthik, S. (eds) Universal Threats in Expert Applications and Solutions. UNI-TEAS 2025

[5] Dr. DarshanabenDipakkumarPandya, Dr. AbhijeetsinhJadeja, Dr. SheshangDegadwala, " An Applied Variation Anomaly Technique for Detection of Irregular Values in Data Mining, IInternational Journal of Scientific Research in Computer Science, Engineering and Information Technology(IJSRCSEIT), ISSN : 2456-3307, Volume 8, Issue 2, pp.143-148, March-April-2022. Available at doi :https://doi.org/10.32628/CSEIT228228.

[6] Gora KanijfatmaMaisurali, HabibaVajiraliMomin, Dr. DarshanabenDipakkumarPandya, Dr. AbhijeetsinhJadeja, " Smart Cook : An AI and Machine Learning Powered Culinary, IInternational Journal of Scientific Research in Computer Science, Engineering and Information Technology(IJSRCSEIT), ISSN : 2456-3307, Volume 10, Issue 1, pp.55-59, January-February-2024.

[7] YashNaraynbhai Patel, Dr. DarshanabenDipakkumarPandya, Dr. AbhijeetsinhJadeja, " An Influence of Ethical Hacking in Civilization , International Journal of Scientific Research in Science and Technology(IJSRST), Online ISSN : 2395-602X, Print ISSN : 2395-6011, Volume 10, Issue 1, pp.195-200, January-February-2023. Available at doi : **https://doi.org/10.32628/IJSRST2310119**

[8] Pandya, D., Jadeja, A., Bhuptani, M., Patel, V., Mehta, K., &Brahmbhatt, D. (2024). Machine Learning: Enhancing Cyber-security through Attack Detection and Identification. *ITM Web of Conferences, 65*, 03010. https://doi.org/10.1051/itmconf/20246503010

[9] Pandya, D., Jadeja, A., Khan, M.A., Trivedi, S.B., Ramnath, M.A., Satish, B.P. (2024). Significance of Sentiment Analysis with Text-Based Mining Approach. In: Rathore, V.S., Piuri, V., Babo, R., Tiwari, V. (eds) Emerging Trends in Expert Applications and Security. ICE-TEAS 2024. Lecture Notes in Networks and Systems, vol 1037. Springer, Singapore. https://doi.org/10.1007/978-981-97-3991-2_27

[10] HeemaPatadiya, Akshay Patel, FalguniRahevar, and Dr. DarshanabenDipakkumarPandya, "Revolutionizing Banking Industry : The Role of Fintech in Modern Banking Industry", *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol*, vol. 11, no. 4, pp. 20–25, Jul. 2025, doi: 10.32628/CSEIT25113448.