Determination of the Best Imputation Technique for Estimating Missing Values when Fitting the Weibull Distribution

Ahmad Shukri Yahaya
Nor Azam Ramli
Fauziah Ahmad

Clean Air Research Group
School of Civil Engineering
Universiti Sains Malaysia
Engineering Campus, 14300 Nibong Tebal
Seberang Perai Selatan
PULAU PINANG.

Norlida Mohd Nor
Muhammad Nor Hakim Bahrim

School of Mathematical Sciences
Universiti Sains Malysia
11800 Minden, PULAU PINANG

Abstract

Missing values are common in any scientific work. Nine single imputation techniques were compared to determine the best technique for estimating missing values when fitting Weibull distributions. Simulation technique was used to obtain random variables for the Weibull distributions. Samples of sizes 30, 50, 200 and 300 were used to simulate the Weibull distribution. To determine the best imputation technique, four error measures were used that are the normalized absolute error, root mean square error, index of agreement and root mean square error. This study shows that no single imputation technique is the best for each sample size and for each percentage of missing values.

Keywords Weibull distribution, Imputation techniques, Performance indicators

1. INTRODUCTION

Missing value is very common in many scientific fields such as statistical, clinical, psychology, environmental research and others. In environmental research especially in air pollution modelling, observations are usually collected by using specialized machines to obtain air pollutants concentrations, wind speed, wind direction, rainfall amount and temperature. These machines sometimes failed to register the observations due to machine malfunction and sometimes when the machine needs servicing. This will result in the occurrence of missing values. Missing data hinder the ability to make exact conclusion or interpretations about the observation. Thus various missing value techniques have been developed to overcome this problem.

Little and Rubin (2002) has discussed about three classifications of missing data that are missing completely at random (MCAR), missing at random (MAR) and non-ignorable missing data (NMAR). Missing completely at random (MCAR) means that the missing value does not depend on any data or variable either observed or missing. The missing data occurs randomly in the datasets. Meanwhile, data are said to be missing at random (MAR) if the failure to observe a value does not depend on the value that would be observed. Non-ignorable missing data (NMAR) occur when missing values are not randomly distributed across observations. For MCAR and MAR, missing values can be deleted. Two common procedures for deletion of missing data are listwise deletion and pairwise deletion. Little and Rubin (2002) indicates that listwise deletion is not an appropriate way to handle missing data because by deleting the data, it will decrease the sample size and resulting in the decreasing of statistical power. Meanwhile the pairwise deletion may preserve more information. Thus imputation of missing values should be considered. The use of single imputation techniques for estimating missing values when fitting the Weibull distribution for carbon monoxide data was discussed by Yahaya et al. (2005).

They used the mean-based imputation technique and found that the mean estimation method using one datum above and one datum below is the best method. The percentage of missing values is about seven percent. The mean-based imputation technique was used by Yahaya et al. (2008) to estimate the missing values. They used simulation techniques to obtain random observations from five different Weibull distributions that represent left-skewed distributions, normal distribution and right-skewed distributions. The percentage of missing values used was 5%. This simulation study shows that the mean all method is the best imputation method to be used. Mohamed Noor Norazian (2008) describes the use of single imputation techniques to estimate missing values for a one year hourly PM_{10} data. They used the linear quadratic and cubic interpolation techniques, nearest neighbour, mean-before-after and mean-before as the imputation techniques. Simulated percentage missing values of 5%, 10%, 15%, 25% and 40% were used. They concluded that the best imputation technique is mean-before-after. Junninen et al. (2004) describes in detail various single and multiple imputation techniques that can be used for six air quality datasets. They found that the more complex imputation techniques known as hybrid models and multiple imputations are better than the other techniques.

Plaia and Bondi (2006) analyzed the space-time variability of PM_{10} concentration using meteorological variables. This paper uses space-time information on PM_{10} level concentrations in eight monitoring sites that are called Site-Dependent Effect method (SDEM). Four simulated incomplete data were generated and five methods including SDEM have been applied. They found that the best imputation techniques are those that consider both space and time information that is the SEDEM method. Weibull distribution have been used to fit distributions in air pollution studies and to determine return periods (Yusoff et al., 2009; Seinfeld and Pandis, 1997; Maffeis ,1998). This distribution have also been used successfully in fitting distributions for wind speed (Shoji , 2005; Jaramillo and Borja,2004 and Yahaya et al. (2007). Weibull distribution has also been used in life testing and reliability theory. This paper compares nine single imputation methods for estimating missing values for the Weibull distributions of sizes 30, 50, 200 and 300. Five percentages of missing values were chosen at random that are 5%, 10%, 15%, 20% and 25%. These imputation techniques were chosen because it can be easily computed.

2. MATERIALS AND METHODS

2.1 The Weibull distribution

This distribution was originally derived by Fisher and Tippet in 1928 as an asymptotic extreme value distribution. In 1939 the Swedish physicist Weibull derived the same distribution on the basis of practical requirements in the analysis of material breaking strength. It was not until 1951, however, when one of Weibull's articles received wide circulation among engineers concerned with modelling the statistical variation of their data, that this distribution became prominent in the engineering community. Weibull's name has since been associated with this distribution (Bury, 1999). The Weibull density function contains two parameters, sigma (σ) and mu (μ). The σ value acts as a scale parameter and μ value acts as the location parameter that determines the form and 'skew' of the distribution (Piegorsch and Bailer, 1997).

$$f(x)=\left(\frac{\lambda}{\sigma}\right)\left(\frac{x}{\sigma}\right)^{\lambda-1}\exp\left\{-\left(\frac{x}{\sigma}\right)^{\lambda}\right\} \qquad x\geq 0, \quad \sigma,\lambda\geq 0 \qquad (1)$$

The cumulative distribution function (cdf) takes the form

$$F(x)=1-\exp\left\{-\left(\frac{x}{\sigma}\right)^{\lambda}\right\} \qquad x\geq 0, \quad \sigma,\lambda\geq 0 \qquad (2)$$

where σ is the scale parameter and λ is the shape parameter. The scale parameter controls the spread of the distribution, and the shape parameter controls the form of the distribution.

2.1.2 Single imputation methods

Nine imputation methods were used to deal with missing values that are mean above, mean below, mean above below, median above and median below, nearest neighbour, linear interpolation, spline interpolation and regression (Yahaya *et al.*, 2005). The description of method used in this study is provided in Table 1.

Let $y_1, y_{2,\ldots,} y_n$ be the *n* observations of a time series data and *k* are the missing values that are denoted by

$y_1^*, y_{2,\ldots,}^* y_k^*$ with $k < n$. Than the observed data with missing values are given by

$$y_{1,} y_{2,} \ldots, y_{n1,} y_1^*, y_{n1+1}, y_{n1+2,\ldots,} y_{n2}, y_{2,}^* y_{n2+1}, y_{n2+2,\ldots,} y_k^*, y_n$$

The formulae for the first estimated values using the imputation methods are given in Table 1 according to the notations given above.

Table 1 Single imputation methods

2.2 Performance Measures

Four types of indicators were used to verify the best plotting formula to estimate the parameters of the Frechet distribution. Two error measures that is the normalized absolute error (*NAE*) and the root mean square error (*RMSE*) and two accuracy measures that is index of agreement (*IA*) and coefficient of determination (R^2) were used. To produce a good estimator, the error measures must approach zero and the accuracy measure should approach one. Table 2 gives the formula for the performance indicators.

Table 2 Performance Indicators

3. DATA

Simulation of random variables for Weibull distribution was used to obtain the observed values. The values of the shape and scale parameters were chosen to represent various forms of the Weibull distribution. The range of values for the shape parameter is $(0,5)$ and for the scale parameter it is from 20 to 90. These values were chosen because it represents the Weibull distribution that is often used in air pollution studies in Malaysia (Sedek et al., 2006; Yusoff et al., 2009; Nurulilyana Sansuddin, 2010).

Sample of sizes 30, 50, 100, 200 and 300 were used. Each of the sample sizes were replicated ten times for each Weibull distribution. For each sample sizes, 100000 random variables were generated. Multiplicative congruential generator (Law and Kelton, 2000) of the form

$$X_i = (397204094 \ X_{i-1}) \bmod (2^{31} - 1)$$

was used to simulate the random variables. The above multiplicative congruential generator was found to have good statistical properties.

From these simulated Weibull random variables 5%, 10%, 15%, 20% and 30% missing values were created at random. Air pollution data are usually obtained using automated machines and it was found that usually about less 20% of data are missing (Sedek et al., 2006; Yusoff et al., 2009; Nurulilyana Sansuddin, 2009). These missing values were than estimated using the eleven imputation methods. The errors between observed and estimated missing values were then obtained and the best imputation method was obtained.

4. RESULTS AND DISCUSSIONS

The results of the simulation study using mean-based imputation, median-based imputation, nearest neighbour, linear interpolation, cubic spline and linear regression are given in this section. Table 3 gives the result of the performance indicators when there are 5% missing values. When sample size is 30, the best technique is nearest neighbour and followed by median above and median below respectively. For this case, only the normalized absolute error was obtained because the number of missing value is very small. When $n = 50$ median below is the best and followed by median above and nearest neighbour. The cubic spline technique dominates for $n = 200$. The median based technique is the next best imputation technique. However, nearest neighbour is the top and followed by linear interpolation and median above technique. Overall when the missing value is 5%, the best imputation techniques are nearest neighbour and the two median-based techniques.

Table 3 Performance Indicators with 5% missing values

Table 4 shows the performance of the imputation techniques when the percentage of missing value is 10%. When the sample size is 30, linear regression technique is found to be the best imputation technique. However, for sample sizes 50, 200 and 300, the nearest neighbour is the best. Thus when there are 10% of missing values, the technique to be used is the nearest neighbour and this is followed by linear interpolation and mean above.

Table 4 Performance Indicators with 10% missing values

The results of the imputation techniques for 15% missing values are given in Table 5. For sample of size 30, median below gives the best result and for sample size 200, the mean above is the best imputation technique. The best imputation technique for sample sizes 50 and 300 is the nearest neighbour. The best techniques to be considered when there are 15% missing values are linear interpolation, nearest neighbour and mean above.

Table 5 Performance Indicators with 15% missing values

Table 6 shows the performance measures at 20% missing values for all studied sample sizes. The best imputation technique for $n = 30$ and $n = 50$ is linear interpolation. For $n = 200$, the best imputation technique is the mean below whereas for $n = 300$ median below is the best. This result show that the best imputation technique varies according to sample size. However without considering the percentage of missing values, the best techniques are linear interpolation, mean above and mean below. Median below is at ranking number four.

Table 6 Performance Indicators with 20% missing values

At 25% missing values, the results of the performance indicators for the imputation techniques are shown in Table 7. Again the best imputation techniques depend on the sample sizes. Nearest neighbour technique is shown to be the best for $n = 30$ and $n = 50$ and is second best for $n = 200$. Median above is best for $n = 200$ and second best for $n = 30$ and $n = 50$. Linear interpolation technique is shown to be the best for $n = 300$. At 25% missing values, the best technique is nearest neighbour, median below and mean below.

Table 7 Performance Indicators with 25% missing values

We also look the best imputation techniques for each sample size. When the sample sizes are small ($n = 30$ and $n = 50$), the best techniques are median below and nearest neighbour. For $n = 200$, the best techniques are median below and median above. For $n = 300$, the best techniques are linear interpolation and nearest neighbour with the median above being the fourth best technique.

5. CONCLUSION

Estimating missing values are very important in order to obtain an accurate representation of the data. This paper compares nine simple imputation techniques for estimating missing values when fitting the Weibull distribution. Weibull distribution was chosen because it is widely used in air pollution studies, hydrologic studies and reliability studies. The data for the study was simulated so that it represents air pollutants concentration levels. Sample of sizes 30 and 50 were used to represent small sample and with 200 and 300 to represent large sample. The percentage of missing values used for this study are 5% representing small percentage of missing values, 10% and 15% representing medium percentage of missing values , 20% and 25% representing large percentage of missing values. The results of the study show that no single imputation technique is the best for each sample size and for each percentage of missing values. For small sample sizes, the best techniques are median above, median below and nearest neighbour and for large sample sizes, the best techniques are medium above, linear interpolation and nearest neighbour. This shows that resistant estimators give favourable results because the simulated Weibull distributions are skewed to the right and thus contains extreme values. For small percentage of missing values, the imputation technique that should be considered are nearest neighbour, median above and median below. In contrast, for medium and large percentage of missing values the techniques that should be used are linear interpolation and nearest neighbour.

6. REFERENCES

Bury, K. (1999). *Statistical Distribution in Engineering.* New York: Cambridge University Press

Jaramillo, O.A. and Borja, M.A.(2004). Wind speed analysis in La Ventosa, Mexico: a bimodal probability distribution case. *Renewable Energy*, 29, 1613–1630

Junninen, H., Niska, N., Tuppurainen, K., Ruuskanen, J. & Kolehmainen, M. (2004). Methods for Imputation for Missing Values in Air Quality Data Sets. *Atmospheric Environment*, 38, 2895-2907

Little, R.J.A. & Rubin, D.B. (2002) *Statistical Analysis with Missing Data*. 2nd edition. Canada : John Wiley and Sons

Maffeis, G. (1998) .Prediction of carbon monoxide acute air pollution episodes. Model formulation and first application in Lombardy. *Atmospheric Environment*, 33 (23), 3859 – 3872

Mohamed Noor Norazian, Yahaya Ahmad Shukri, Ramli Nor Azam, Abdullah Mohd Mustafa Al Bakri (2008). Estimation of missing values in air pollution data using single imputation techniques, *ScienceAsia Journal*, 34, 341-345

Nurulilyana Sansuddin (2009) *Modeling Locational Differences And Prediction Of TemporalConcentration Of PM_{10} Using Time Series Analysis*, Ph.D Thesis, University Sains Malaysi

Piegorsch, W. W. and Bailer, A. J. (1997) *Statistic for Environment Biology and Toxicology*. London: Chapman & Hall Publication

Plaia, A. and Bondi, A.L. (2006) Single Imputation Method of Missing Values in Environmental Pollution Data Sets. *Journal of Atmospheric Environment*, 40, 7316-7330

Sedek, J. N. M., Ramli, N. A. and Yahaya, A. S. (2006) Air quality predictions using lognormal distribution functions of particulate matter in Kuala Lumpur, Malaysia. *Journal of Environmental Management*, 7, 33 – 41

Seinfeld, J. H. and Pandis, S. N. (1997). *Atmospheric Chemistry and Physics, from Air Pollution to Climate Change*, New York: Wiley

Shoji, T. (2005). Statistical and geo-statistical analysis of wind: A case study of direction statistics, *Computers & Geosciences*, 32, 1025-1039

Yahaya, A.S., Ramli, N.A, and Yusof, N.F.F.M. (2005) Effect of estimating missing values on fitting distributions, *Proceeding of The International Conference on Qualitative Sciences and Its Applications (ICOQSIA)*, 1-5. 6-8 December, Penang, Malaysia

Yahaya, A.S., Ramli, N.A., Abdul Wahab, A.A.(2007). Finding The Best Wind Speed Distribution: A Case Study, *World Engineering Congress 2007 (WEC2007)*, 5-9 August 2007, Penang, Malaysia, 14-19.(CD Proceedings

Yahaya, A.S., Abdul Hamid, H., Mohd Nor, N., Wong, L. S. (2008) Comparison Of Mean Imputation Techniques In Estimating Missing Values For The Weibull Distribution, *The Poceedings of the 16th National Mathematical Sciences Symposium, Symposium*, 2-5 June 2008, Kota Bharu, Kelantan, 183-188

Yusoff, N.F.F., Ramli, N.A., Yahaya, A.S., Sansuddin, N., Ghazali, N.A., AlMadhoun, W.A. (2009). Monsoonal Differences and Probability Distribution of PM10 Concentration. *Environmental Monitoring & Assessment 163(1-4)*, 655-667

6. TABLES AND FIGURES

Table 1 Single imputation methods

Methods	Description and Formula
Mean Above - The missing values will be replaced by the mean value of the above missing data	$y_1^* = \frac{1}{n_1}\sum_{i=1}^{n_1} y_i$
Mean Below- The missing values will be replaced by the mean value of the below missing data	$y_1^* = \frac{1}{(n_2 - n_1 + 1)}\sum_{i=n_1+1}^{n_2} y_i$
Mean Above Below - The missing values will be replaced by the average of one existing data above and below the missing values	$y_1^* = \frac{1}{2}\left(y_{n_1} + y_{n_1+1}\right)$ (for data ranked from smallest to largest)
Median Above - The missing value will be replaced by the median value of the above missing data	$y_1^* = \begin{cases} y\left(\frac{n_1+1}{2}\right) & n \text{ odd} \\ \frac{y\left(\frac{n_1}{2}\right) + y\left(\frac{n_1}{2}+1\right)}{2} & n \text{ even} \end{cases}$ (for data ranked from smallest to largest)
Median Below - The missing value will be replaced by the median value of the below missing data	$y_1^* = \begin{cases} y\left(\frac{n_2-n_1+1}{2}\right) & n \text{ odd} \\ \frac{y\left(\frac{n_2-n_1+1}{2}\right) + y\left(\frac{n_2}{2}+1\right)}{2} & n \text{ even} \end{cases}$ (for data ranked from smallest to largest)
Linear Interpolation - The missing value will be replaced by drawing a straight line between two neighbouring data	$y_i^* = y_0 + (y_1 - y_0)\left(\frac{x_i - x_0}{x_1 - x_0}\right)$ where $\left(x_0, y_0\right)$ and $\left(x_1, y_1\right)$ are the initial and last values
Spline Interpolation - The missing value will be replaced by using low degree polynomials in each of the interval point	$y_i^* = a_0 + a_1 x + a_2 x^2 + a_3 x^3$ where $x_k < x < x_{k+1}$
Nearest Neighbour - The missing value will be replaced by the nearest value	$y_i^* = y_{i-1}$ or y_{i+1}
Linear Regression - The missing value will be replaced by regression of the unobserved variables against observed ones for that datasets	$y_i^* = \beta_0 + \beta_1 x_i + \varepsilon_i$ where β_0 and β_1 are the intercept and slope parameters of the regression model and are estimated by the least squares method, ε_i is the error term

Table 2 Performance Indicators

Measures	Formula
Normalized absolute error (*NAE*)	$NAE = \frac{\sum_{i=1}^{N} \lvert P_i - O_i \rvert}{\sum_{i=1}^{N} O_i}$
Root Mean Square Error (*RMSE*)	$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (P_i - O_i)^2}$
Index of Agreement (*IA*)	$1 - \left[\frac{\sum_{i=1}^{N} (P - O_i)^2}{\sum_{i=1}^{N} \left(\lvert P_i - \bar{O} \rvert + \lvert O_i - \bar{O} \rvert \right)^2} \right]$
Coefficient of Determination(R^2)	$R^2 = \left[\frac{1}{N} \frac{\sum_{i=1}^{N} \left[\left(P_i - \bar{P} \right) + \left(O_i - \bar{O} \right) \right]}{\sigma_P \sigma_O} \right]^2$

*N is the number of observations, O_i is the observed values , P_i is the predicted values, $\bar{O}$ is the average of the observed values, $\bar{P}$ is average of predicted values, σ_P is standard deviation of observed values and σ_O is standard deviation of observed values.

Table 3 Performance Indicators with 5% missing values

Imputation Method	NAE				RMSE				R^2				IA			
Sample size →	30	50	200	300	30	50	200	300	30	50	200	300	30	50	200	300
Nearest	28.935	3.305	2.357	0.382	NA	65.131	106.81	43.310	NA	0.25	0.690	0.767	NA	0.509	0.675	0.903
Mean Abv	199.74	4.896	1.544	0.596	NA	68.396	72.174	55.616	NA	0.25	0.597	0.658	NA	0.386	0.641	0.746
Mean AB	113.95	4.162	1.456	0.681	NA	59.063	67.751	61.515	NA	0.25	0.710	0.664	NA	0.377	0.606	0.643
Mean Bel	47.133	3.481	1.307	0.580	NA	50.715	65.697	55.323	NA	0.25	0.578	0.665	NA	0.439	0.654	0.746
Med Abv	100.39	2.674	1.038	0.557	NA	53.131	66.445	61.290	NA	0.25	0.636	0.762	NA	0.475	0.679	0.723
Med Bel	36.507	1.834	0.778	0.559	NA	45.997	58.261	61.597	NA	0.25	0.615	0.760	NA	0.442	0.693	0.721
Linear	226.07	5.311	1.783	0.434	NA	69.721	80.784	45.980	NA	0.25	0.668	0.726	NA	0.423	0.688	0.869
Spline	252.82	6.982	0.813	0.635	NA	91.472	58.047	53.158	NA	0.25	0.621	0.666	NA	0.401	0.727	0.897
Regression	176.52	5.781	1.339	0.987	NA	67.952	70.740	81.700	NA	0.25	0.614	0.485	NA	0.255	0.405	0.235

Table 4 Performance Indicators with 10% missing values

Imputation Method	NAE				RMSE				R^2				IA			
Sample size →	30	50	200	300	30	50	200	300	30	50	200	300	30	50	200	300
Nearest	3.132	0.654	0.648	0.551	236.069	67.240	68.368	64.561	0.336	0.548	0.789	0.808	0.383	0.725	0.807	0.842
Mean Abv	1.469	0.774	0.590	0.591	167.283	76.354	67.195	71.037	0.325	0.422	0.768	0.703	0.529	0.578	0.750	0.724
Mean AB	1.512	0.869	0.656	0.601	165.724	81.462	73.701	61.945	0.399	0.409	0.723	0.742	0.500	0.496	0.645	0.711
Mean Bel	1.612	0.841	0.591	0.594	167.996	79.246	67.044	71.086	0.334	0.393	0.762	0.704	0.494	0.544	0.749	0.723
Med Abv	1.239	0.659	0.664	0.552	167.328	77.109	80.210	76.192	0.347	0.468	0.738	0.713	0.582	0.570	0.632	0.689
Med Bel	1.149	0.697	0.652	0.559	163.136	79.263	79.180	76.392	0.367	0.508	0.751	0.714	0.576	0.564	0.634	0.688
Linear	2.239	0.539	0.651	0.504	179.713	76.127	74.443	53.789	0.432	0.407	0.812	0.814	0.479	0.557	0.791	0.815
Spline	4.458	0.879	1.112	0.734	255.631	73.599	102.07	66.514	0.413	0.468	0.780	0.770	0.447	0.724	0.787	0.827
Regression	1.221	1.040	0.881	0.873	151.649	89.110	87.761	73.526	0.381	0.482	0.601	0.598	0.581	0.339	0.341	0.296

Table 5 Performance Indicators with 15% missing values

Imputation Method	NAE				RMSE				R^2				IA			
Sample size →	30	50	200	300	30	50	200	300	30	50	200	300	30	50	200	300
Nearest	1.169	0.647	0.431	0.291	49.242	66.273	56.308	36.781	0.457	0.561	0.789	0.845	0.708	0.703	0.843	0.933
Mean Abv	0.846	0.647	0.406	0.374	45.419	66.541	37.888	39.975	0.340	0.576	0.794	0.821	0.685	0.691	0.838	0.902
Mean AB	0.824	0.629	0.495	0.464	42.756	66.559	41.785	45.591	0.405	0.572	0.797	0.850	0.588	0.647	0.771	0.853
Mean Bel	0.803	0.686	0.412	0.381	41.284	69.469	37.937	40.183	0.355	0.587	0.790	0.819	0.672	0.630	0.836	0.901
Med Abv	0.708	0.597	0.407	0.324	44.989	76.348	40.542	41.570	0.403	0.470	0.838	0.809	0.677	0.679	0.833	0.879
Med Bel	0.644	0.714	0.403	0.328	40.382	81.046	40.217	41.881	0.405	0.525	0.835	0.804	0.713	0.603	0.832	0.877
Linear	0.781	0.670	0.404	0.320	44.111	68.159	46.980	37.693	0.420	0.639	0.778	0.868	0.776	0.691	0.820	0.926
Spline	2.976	1.286	0.842	0.600	133.46	118.86	85.263	53.234	0.408	0.580	0.561	0.791	0.563	0.646	0.695	0.918
Regression	1.003	0.818	0.821	0.869	49.367	75.826	61.350	82.299	0.425	0.557	0.716	0.556	0.559	0.486	0.297	0.240

Table 6 Performance Indicators with 20% missing values

Imputation Method	NAE				RMSE				R^2				IA			
Sample size →	30	50	200	300	30	50	200	300	30	50	200	300	30	50	200	300
Nearest	0.842	0.669	0.762	0.282	93.336	86.572	54.262	34.745	0.589	0.665	0.898	0.855	0.764	0.646	0.836	0.939
Mean Abv	0.836	0.626	0.675	0.335	92.615	78.430	42.365	40.489	0.577	0.688	0.847	0.867	0.746	0.721	0.848	0.899
Mean AB	0.840	0.606	0.725	0.444	95.791	83.802	41.551	48.144	0.487	0.631	0.847	0.860	0.678	0.640	0.814	0.812
Mean Bel	0.790	0.523	0.674	0.333	94.390	74.823	42.243	40.919	0.517	0.632	0.846	0.867	0.749	0.729	0.848	0.894
Med Abv	0.765	0.625	0.554	0.351	91.355	84.367	41.885	42.572	0.541	0.638	0.868	0.867	0.730	0.629	0.840	0.884
Med Bel	0.672	0.584	0.543	0.358	89.158	82.811	41.699	43.621	0.530	0.634	0.867	0.866	0.763	0.636	0.840	0.874
Linear	0.658	0.540	0.640	0.327	86.612	79.923	43.678	40.915	0.566	0.660	0.868	0.870	0.768	0.691	0.865	0.899
Spline	1.331	1.170	1.081	0.702	107.85	131.02	78.172	67.949	0.547	0.608	0.854	0.576	0.696	0.571	0.782	0.708
Regression	1.092	0.871	0.958	0.855	114.10	93.130	55.757	80.284	0.487	0.484	0.661	0.538	0.444	0.513	0.343	0.164

Table 7 Performance Indicators with 25% missing values

Imputation Method	NAE				RMSE				R^2				IA			
Sample size →	30	50	200	300	30	50	200	300	30	50	200	300	30	50	200	300
Nearest	0.529	0.362	0.411	0.406	38.119	32.015	33.582	42.351	0.608	0.765	0.883	0.865	0.844	0.939	0.931	0.907
Mean Abv	0.698	0.379	0.420	0.369	46.986	36.525	33.555	36.848	0.445	0.705	0.881	0.873	0.657	0.899	0.916	0.932
Mean AB	0.620	0.441	0.529	0.449	43.540	37.973	38.360	37.269	0.633	0.701	0.839	0.876	0.794	0.838	0.869	0.899
Mean Bel	0.594	0.387	0.417	0.363	42.708	37.264	33.551	36.658	0.551	0.686	0.877	0.871	0.812	0.881	0.917	0.933
Med Abv	0.523	0.305	0.385	0.338	44.187	32.927	33.403	38.649	0.612	0.740	0.886	0.857	0.816	0.910	0.906	0.924
Med Bel	0.498	0.366	0.392	0.333	41.579	40.554	33.698	38.646	0.600	0.687	0.885	0.856	0.829	0.862	0.906	0.924
Linear	0.533	0.418	0.468	0.402	42.789	37.360	33.854	34.830	0.621	0.763	0.865	0.896	0.766	0.889	0.908	0.929
Spline	1.562	0.436	0.949	0.767	110.17	32.996	74.902	78.048	0.553	0.768	0.730	0.791	0.605	0.934	0.802	0.835
Regression	0.993	0.898	0.906	0.970	56.881	71.895	64.117	72.589	0.544	0.560	0.681	0.648	0.567	0.383	0.355	0.197