# Optimal Parameter Choice for Imputing Missing Values in Water Level Data Using the K-Nearest Neighbour (Knn) Method

Nura Umar and Alison Gray

University of Strathclyde, Department of Mathematics and Statistics,
nura.umar@strath.ac.uk; a.j.gray@strath.ac.uk.

Keywords: kNN, missing values, water level, parameter choice, single imputation

Missing values in data is a problem which data analysts in most areas of research must deal with, before analysis. Ignoring missing data values will cause loss of information and efficiency, and unreliable results, especially where large proportions of the dataset are missing [1]. Various approaches are used to manage missing data, notably single and multiple imputation [2]. Imputation methods replace missing data values. Single imputation replaces each missing value with a single value [3], while multiple imputation generates two or more values for each missing value [4]. The k-nearest neighbour (kNN) method for imputation is a single imputation method. It uses the k most similar observations (nearest neighbours) in the dataset to identify a replacement for a given missing value, but a suitable value of k must be chosen beforehand. When k=1, the replacement value is obtained as the observed value for the variable of interest from the nearest case to the case with a missing value. When k>1, the average of k observed values is used as the replacement value, where these observed values are the responses from the k nearest cases. First of all, kNN identifies the k nearest observations to each case with a missing value, by calculating the distances between the observed variables for that case and all other cases in the dataset [5,6].

It is important to identify the most suitable value for k, to achieve the most accurate data replacements. Muinonen et al. [7] state that imputation accuracy from kNN does not improve for k>3, however McRoberts et al. [8] found that a higher value of k (k≥7) may improve estimation accuracy and have lower variability of results. Because of these contrasting views, this work examines the performance of k=1,3,5,7,9,11 and 15, to identify the optimum value of k. It uses monthly water level datasets from the Nigeria Hydrological Services Agency (for years 2011-2016) from three water stations, Ibi, Makurdi and Umaisha water stations on the river Benue in Nigeria. As these data themselves contained missing values, time series models were fitted to each dataset and complete data were simulated from those models for this study. Missing data at rates 10%, 20%, 30%, 40%, and 50% were then created using three missing data mechanisms, missing at random (MAR), missing completely at random (MCAR), and missing not at random (MNAR), on the simulated data and imputation was carried out on these.

Two measures were used to assess performance of kNN with each percentage missing and each value of k, for each water station: root mean square error (RMSE) and mean absolute percentage error (MAPE). The results for the MAR and MNAR missing data mechanisms are similar, suggesting k=7 and k=15 as generally best, but sometimes k=3 also does well, especially for the MCAR mechanism. Overall, k=7 and k=15 consistently produce the best performance (lowest mean and standard deviation values of either RMSE or MAPE) for the five percentages missing and all three missingness patterns considered. In general k=7 is a good choice, but k=15 is better for these datasets. This result agrees with the findings in [8].

These results will allow more accurate replacement of missing values in water level data and hence more reliable conclusions from statistical analysis of these datasets.

References

1. Little, R.J.A. Missing-data adjustments in large surveys. J. Bus. Econ. Stat. 1988, 6, 287-296. https://doi.org/10.2307/1391878

2. Peugh, J.L.; Enders, C.K. Missing data in educational research: a review of reporting practices and suggestions for improvement. Rev. Educ. Res. 2004, 74, 525-556. https://doi.org/10.3102/00346543074004525

3. Enders, C. K. Applied Missing Data Analysis, 1st ed.; Guilford Press: New York, United States, 2010.

4. Graham, J. W.; Hofer, S. M. Multiple imputation in multivariate research. In Modeling longitudinal and multilevel data: Practical issues, applied approaches, and specific examples, Little, T. D.; Schnabel, K. U.; Baumert, J., Eds., Lawrence Erlbaum Associates Publishers: New Jersey, United States, 2000; pp. 201–218, 269-281.

5. Eskelson, B. N.; Temesgen, H.; Lemay, V.; Barrett, T. M.; Crookston, N. L.; Hudak, A. T. The roles of nearest neighbor methods in imputing missing data in forest inventory and monitoring databases. Scandinavian J. Forest Res. 2009, 24, 235-246. https://doi.org/10.1080/02827580902870490

6. Kowarik, A.; Templ, M. Imputation with the R Package VIM. J. Stat. Software 2016, 74, 1-16. https://doi.org/10.18637/jss.v074.i07

7. Muinonen, E.; Maltamo, M.; Hyppänen, H.; Vainikainen, V. Forest stand characteristics estimation using a most similar neighbor approach and image spatial structure information. Remote Sensing Environ. 2001, 78, 223-228. https://doi.org/10.1016/S0034-4257(01)00220-6

8. McRoberts, R. E.; Nelson, M. D.; Wendt, D. G. Stratified estimation of forest area using satellite imagery, inventory data, and the k-nearest neighbors technique. Remote Sensing Environ. 2002, 82, 457-468. https://doi.org/10.1016/S0034-4257(02)00064-0

This is a peer-reviewed, accepted manuscript of the following paper: Umar, N., & Gray, A. (2023). *Optimal parameter choice for imputing missing values in water level data using the k-nearest neighbour (kNN) method*. The Doctoral School Multidisciplinary Symposium (DSMS 2023), Glasgow, United Kingdom.