Imputation Methods and Their Benchmarking for Fuzzy Datasets with R

Maciej Romaniuk

WIT Academy, Poland

Systems Resarch Institute, Polish Academy of Sciences, Poland

mroman@ibspan.waw.pl ORCID: 0000-0001-9649-396X

https://www.researchgate.net/profile/Maciej-Romaniuk

UseR! 2026, Warszawa, 6–9.07.2026

Fuzzy numbers

A fuzzy number [7] is a generalization of a real number (so-called "crisp" value) used to represent imprecise or ambiguous quantities (e.g., to state "the temperature outside is *about* 20°C"). A special *membership function* describes the "degree of truth" that x belongs to the fuzzy set. For the value 1, x is considered fully compatible with the modeled concept, whereas 0 indicates a complete lack of such compatibility. Fuzzy numbers are used to model imprecision, e.g., in engineering, risk analysis, decision-making systems, statistics, etc.

Trapezoidal fuzzy numbers

A trapezoidal fuzzy number $\tilde{a}$ is a fuzzy number defined by four real numbers (a_1, a_2, a_3, a_4) such that $a_1 \leq a_2 \leq a_3 \leq a_4$, while its membership function has a trapezoidal shape.

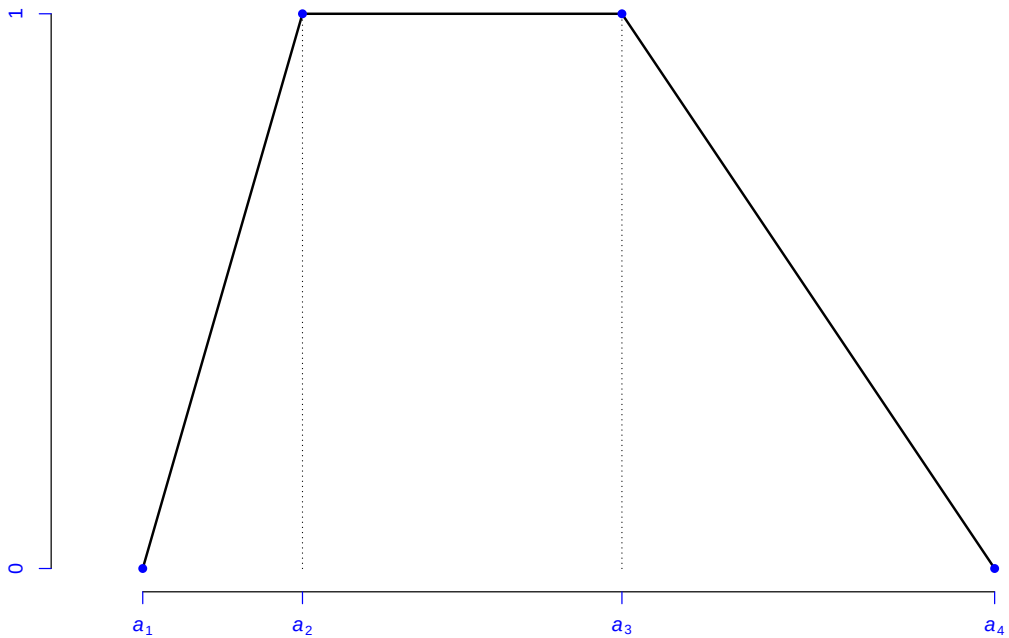


Figure: Example of a trapezoidal fuzzy number

Imputation of dataset

Missing values in datasets pose a significant problem in real-life applications, as they can lead to errors, misleading conclusions, or even false predictions. Statisticians employed various methods to handle missing values (NAs), and imputation (i.e., replacing missing values with appropriately selected substitute data) is among the most important [2].

There are plenty of imputation approaches (e.g., the mean, median, or mode substitution, hot-deck and cold-deck imputation, etc.). Ensuring their quality is also necessary. It can be achieved by applying various benchmarks, which are also part of specialized software [1,5]. However, these imputation methods and benchmarks are designed for "crisp" (i.e., real-valued) datasets.

Imputation of fuzzy-valued dataset

In the fuzzy setting, the problem of missingness is not limited to the availability of a specific value (as in crisp data) but also related to unique features of fuzzy data (such as assumptions about membership functions). Moreover, only the "part" of the whole fuzzy value can be missing, e.g., the left end of its support, and the single adequate real value has to be imputed in such a case.

FuzzyImputationTest package

The R package *FuzzyImputationTest* provides [4,5]:

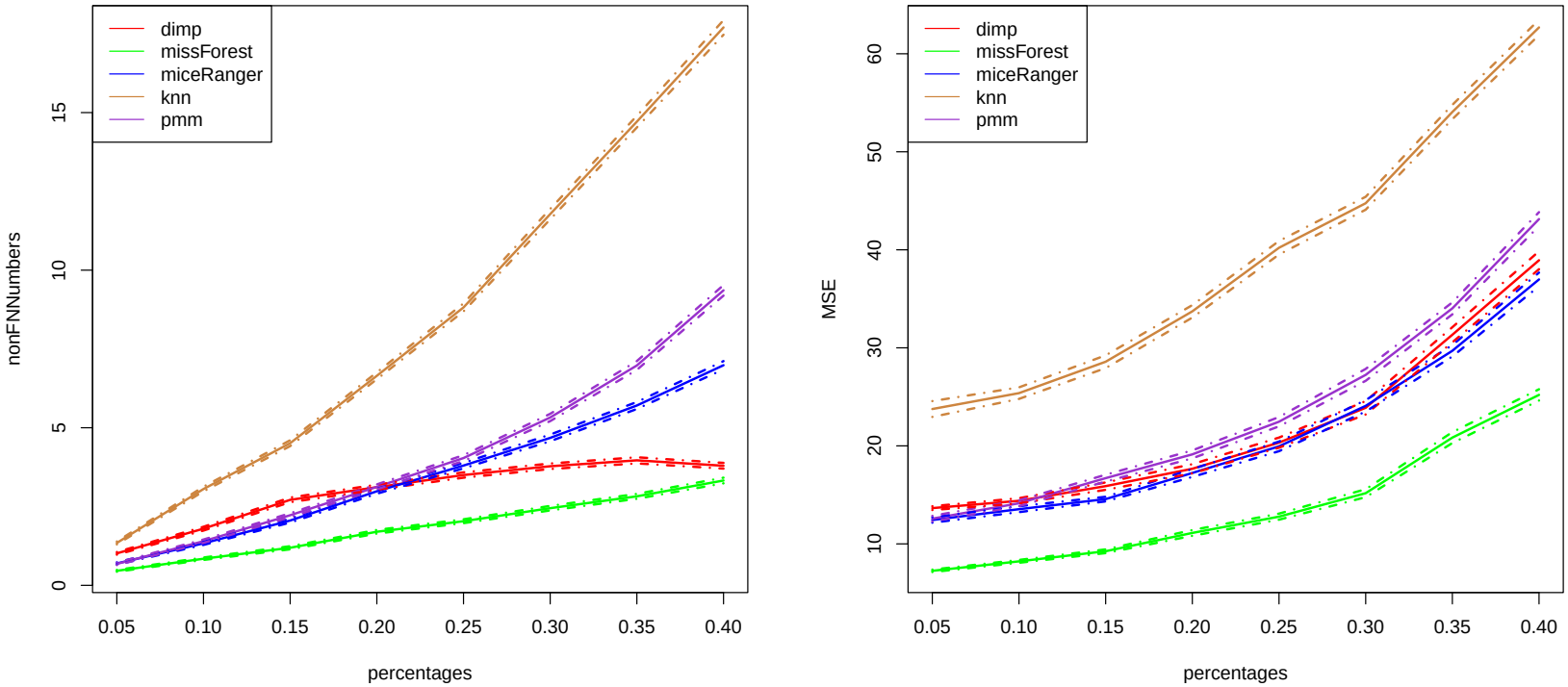
- an extended set of various benchmarks aimed specifically for checking the quality of the imputation methods for fuzzy datasets,
- a user-friendly interface to apply five imputation methods, including widely used ones (*missForest*, *miceRanger*, *knn*, *pmm*, *dimp*) adapted to the fuzzy setting,
- ready-to-use functions for benchmarking imputation algorithms using the given fuzzy dataset.

GamonedoCheese dataset

GamonedoCheese is the dataset that contains experts' opinions concerning their overall impressions of the Gamonedo cheese – a kind of blue cheese produced in Asturias, Spain [3]. These opinions are described by trapezoidal fuzzy numbers.

Using this dataset and *FuzzyImputationTest*, the imputation methods were compared across various percentages of missing values.

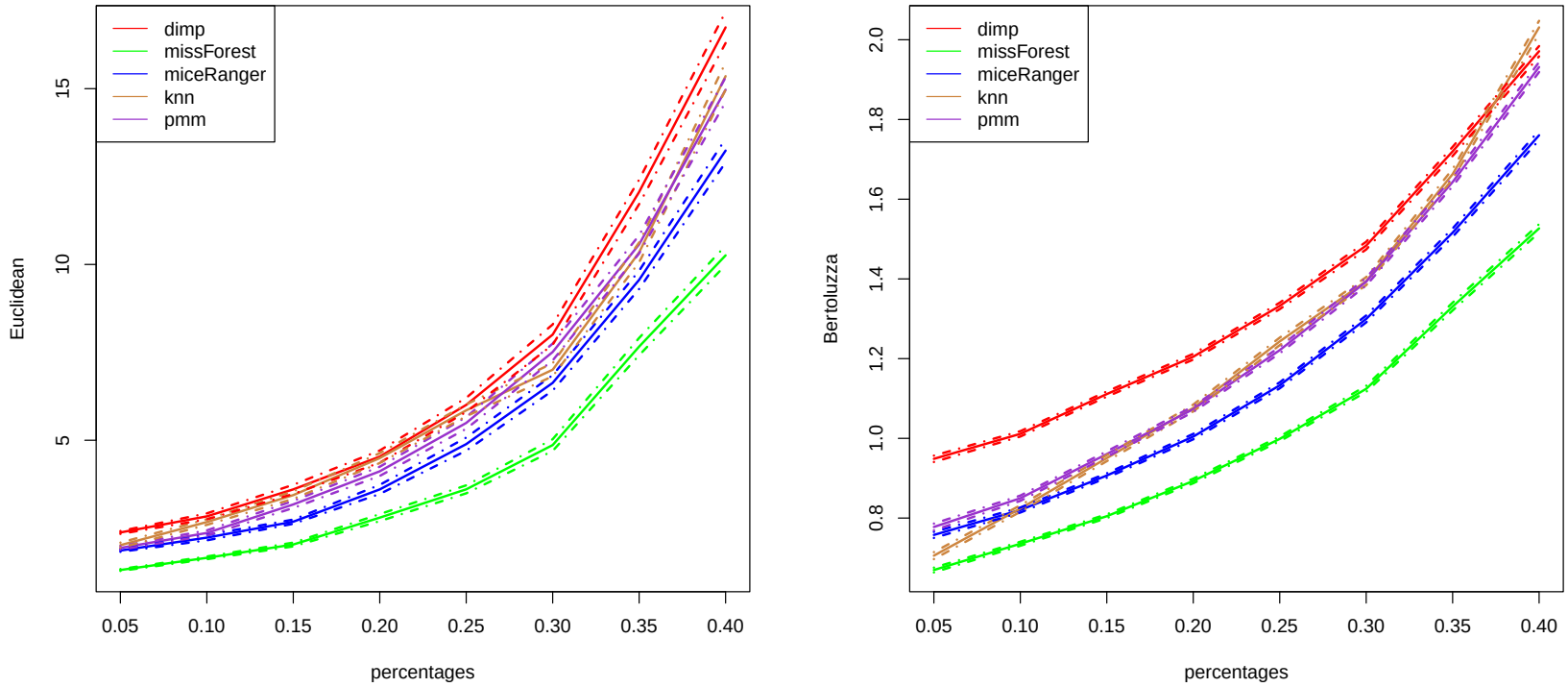
Examples of the outputs



(a) Number of the incorrect FNs (b) MSE error of the imputed values

Figure: Estimated errors as a function of the percentage of missingness p_{imp}.

Examples of the outputs



(a) Euclidean distance

(b) Bertoluzza distance

Figure: Estimated distances between the true and imputed values as a function of the percentage of missingness p_{imp}.

Conclusions

- With the help of the *FuzzyImputationTest* package, five imputation methods were compared for different synthetic, real-life, single- and multivariate fuzzy datasets [4,5].
- From our numerical experiments, it seems that *missForest* should be used primarily for data consisting of only one variable, while *dimp* and *missForest* are advised for multivariate cases.
- *missForest* usually results in the best quality of imputed values, but it may sometimes be slower.
- The behavior of *dimp* is more complex – sometimes it is the best algorithm concerning our benchmarks, and in some cases, the worst one. However, this is the fastest approach.

References

1. Beck M.W. et al., *R Package imputeTestbench to Compare Imputation Methods for Univariate Time Series*, R Journal (2018)
2. Little R., Rubin D., *Statistical Analysis with Missing Data* (2022)
3. Ramos-Guajardo A.B. et al., *Applying Statistical Methods with Imprecise Data to Quality Control in Cheese Manufacturing* (2019)
4. Romaniuk M., *Benchmarking Imputation Methods for Fuzzy Datasets*, International Journal of Applied Mathematics and Computer Science (2026)
5. Romaniuk M., Grzegorzewski P., *Fuzzy Data Imputation with DIMP and FGAIN*, Journal of Computational Science (2026)
6. Yadav M.L., Roychoudhury B., *Handling Missing Values: A Study of Popular Imputation Packages in R*, Knowledge-Based Systems (2018)
7. Zadeh L.A., *Fuzzy Sets*, Information and Control (1965)

Thank you for your attention!