DAMSS 2026
17th International Conference Data Analysis Methods for Software Systems
Druskininkai, Lithuania, 26-28 November 2026

Dataset & code

Data-Driven Imputation of Missing Economic Indicators for EU Regions: A Multi-Metric Accuracy Assessment

Irmantas Pilypas, Irma Šileikienė, Asta Slotkienė Vilnius Gediminas Technical University, Vilnius, Lithuania
irmantas.pilypas@vilniustech.lt, irma.sileikiene@vilniustech.lt, asta.slotkiene@vilniustech.lt

ABSTRACT

Regional economic analysis in the European Union relies on complete statistical data at the NUTS 2 level, yet Eurostat regional datasets frequently contain missing values, particularly for less-developed regions and earlier periods. Mean or median substitution ignores inter-variable relationships, whereas Multiple Imputation by Chained Equations (MICE) preserves them but is computationally demanding for large datasets. Tree-based machine learning methods offer a scalable alternative. However, their accuracy and sensitivity to hyperparameter selection on heterogeneous regional data with varying missingness patterns remain insufficiently studied. This study compares XGBoost and Random Forest for imputing missing regional economic indicators. A dataset of eight economic indicators for 260 regions over the period 1990-2023 was compiled from the Eurostat database (6,984 observations, with 10.01-53.81% missing values per indicator). A separate model was trained for each indicator, with the remaining indicators, region and year as predictors. For each indicator, 20% of the observed values were randomly withheld as a test set. Imputation accuracy was assessed using nRMSE, nMAE, R^2, and sMAPE with five-fold cross-validation on the remaining values for the baseline and hyperparameter-optimised configurations (randomised search over 30 configurations).

RESEARCH AIM

To investigate the differences in the effectiveness of XGBoost and Random Forest for imputing missing values of economic indicators at the EU NUTS 2 level.

DATA

The dataset was compiled from the Eurostat regional database [1]: 260 NUTS 2 regions, eight indicators (ROD1-ROD8), 1990-2023, 6,984 observations. In total, 31.3 % of the indicator values are missing (Table 1), mostly in earlier years and less-developed regions, consistent with a missing-at-random (MAR) mechanism.

Table 1. Economic indicators and share of missing values

ID	Indicator	Missing (%)
ROD1	Employment rate	28.15
ROD2	R&D expenditure (% of GDP)	53.81
ROD3	Unemployment rate	29.04
ROD4	Tertiary education attainment	29.70
ROD5	Youth unemployment rate	36.78
ROD6	Population density	10.01
ROD7	Long-term unemployment share	35.09
ROD8	Female employment rate	28.15

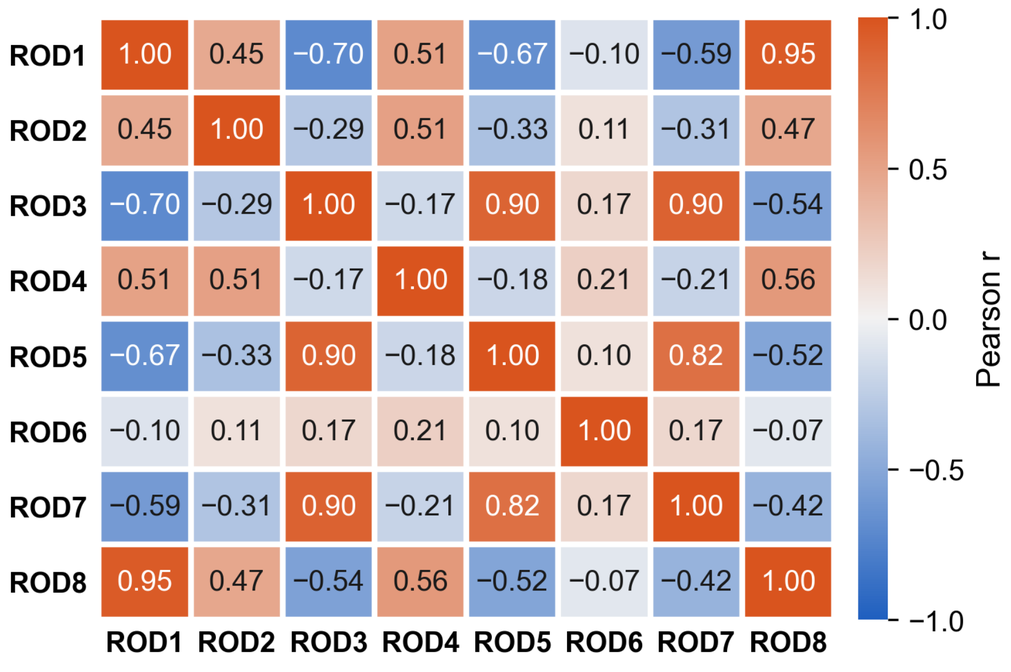

Fig. 1. Pearson correlation matrix of the economic indicators

RESEARCH METHODOLOGY

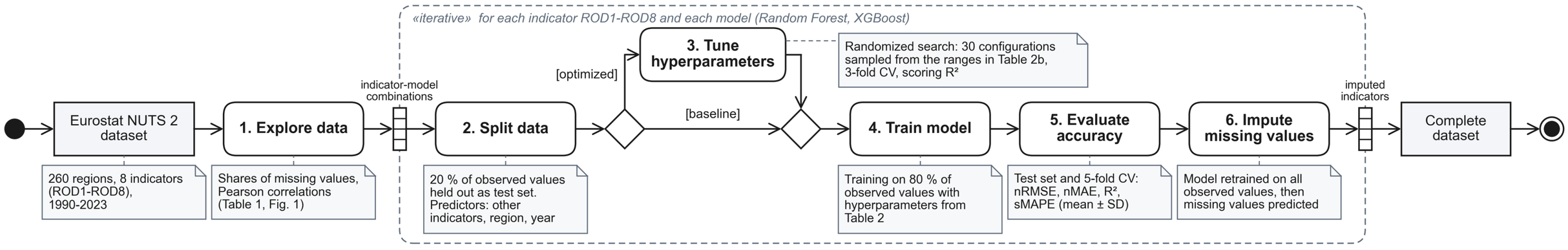

Fig. 2. Research pipeline, executed for each indicator and each model

Table 2a. Baseline configuration - static: fixed values, the same for all indicators

Random Forest	XGBoost
n_estimators = 100	n_estimators = 200
max_depth = 15	max_depth = 6
min_samples_split = 5	learning_rate = 0.1
min_samples_leaf = 2	subsample = colsample_bytree = 0.8
max_features = √p	min_child_weight = 3
-	reg_alpha = reg_lambda = 0.1

Table 2b. Optimized configuration - dynamic: each value drawn uniformly at random from its range or list

Random Forest	XGBoost
n_estimators: integer in 50-300	n_estimators: integer in 50-300
max_depth: 5, 10, 15, 20, 25, 30 or unlimited	max_depth: integer in 3-15
min_samples_split: integer in 2-20	learning_rate: real number in 0.01-0.30
min_samples_leaf: integer in 1-10	subsample, colsample_bytree: real number in 0.6-1.0
max_features: √p, log2 p, 0.3p, 0.5p or 0.7p	min_child_weight: integer in 1-10
-	reg_alpha, reg_lambda: real number in 0-1

RESULTS

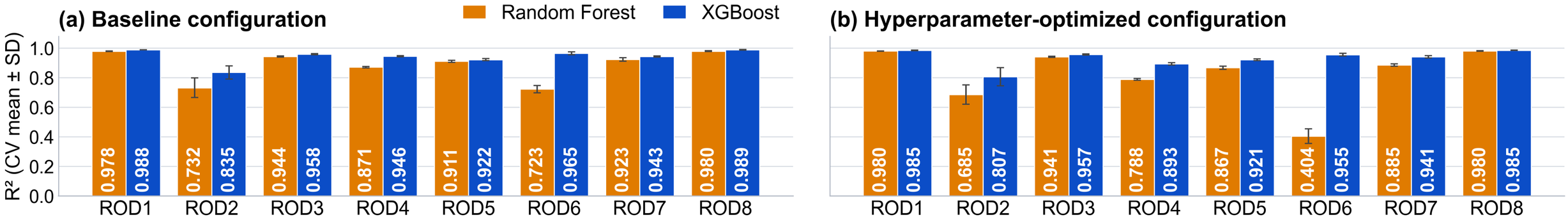

Fig. 3. R^2 by indicator (five-fold CV mean ± SD): (a) baseline, (b) hyperparameter-optimized configuration

CONCLUSIONS

1. XGBoost is the more accurate imputation method: it outperformed Random Forest for all 8 indicators and all 4 metrics (mean R^2 0.9433 vs 0.8826, nRMSE 0.0351 vs 0.0484).

2. Its advantage is largest for indicators weakly correlated with the others (population density: R^2 +33.51 %), and both models exceed $R^2 = 0.97$ for the employment indicators.

3. Hyperparameter optimization did not improve accuracy (mean R^2 −0.066 for Random Forest, −0.013 for XGBoost). Parsimonious baseline configurations are preferable because higher model complexity does not necessarily improve imputation [2] and tuning on small samples risks over-tuning [3].

4. Future work: spatial-temporal dependencies [4], MNAR mechanisms and the effect of imputed-data quality on downstream AI models.

REFERENCES

[1] Eurostat. Regional statistics by NUTS classification (database). https://ec.europa.eu/eurostat/web/regions/database

[2] Hanna, B., Xu, G., Wang, X., Hossain, J. A systematic comparison of machine learning models for missing value imputation in household electricity consumption data. Scientific Reports 16, 26698, 2026. https://doi.org/10.1038/s41598-026-51159-w

[3] Bischl, B., Binder, M., Lang, M., et al. Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges. WIREs Data Mining and Knowledge Discovery 13(2), e1484, 2023. https://doi.org/10.1002/widm.1484

[4] Zahmatkesh, S., Zech, P. Spatio-temporal missing data imputation: A systematic literature review with a focus on statistical and machine learning-based approaches. ACM Computing Surveys 58(10), 1-41, 2026. https://doi.org/10.1145/3797903