PressAcademia
www.pressacademia.org/conferences

IFC
ISBN 978-605-06192-8-7

PressAcademia Procedia YEAR 2023 VOLUME 18

12th Istanbul Finance Congress, December 21, 2023, Istanbul, Turkiye.

HANDLING MISSING VALUES IN MIXED PANEL FINANCIAL DATA: A COMPARISON OF DIFFERENT TECHNIQUES

DOI: 10.17261/Pressacademia.2023.1869
PAP- V.18-2023(27)-p.103-104

Cumhur Ekinci[1], Mustafa Abdullah Hakkoz[2], Unsal Kiran[3], Sirma Seker[4]
[1]Istanbul Technical University, Department of Management Engineering, Istanbul, Turkiye.
ekincicu@itu.edu.tr, ORCID: 0000-0002-0475-2272
[2]Istanbul Technical University, Department of Computer Engineering, Istanbul, Turkiye.
hakkoz22@itu.edu.tr, ORCID: 0000-0002-2963-8513
[3]Istanbul Technical University, Department of Management Engineering, Istanbul, Turkiye.
kiranu20@itu.edu.tr, ORCID: 0000-0003-1813-8748
[4]Istanbul Technical University, Department of Management Engineering, Istanbul, Turkiye.
seker16@itu.edu.tr, ORCID: 0000-0002-2823-9078

To cite this document
Ekinci, C., Hakkoz, M.A., Kiran, U., Seker, S., (2023). Handling missing values in mixed panel data: a comparison of different techniques. PressAcademia Procedia (PAP), V.18, 103-104.
Permanent link to this document: http://doi.org/10.17261/Pressacademia.2023.1869

ABSTRACT

Purpose- The purpose of this study is to compare the success of alternative data imputation techniques with missing data. The study distinguishes itself from the rest of the literature by proposing an appropriate technique for mixed data on financial performance and environmental, social and governance (ESG) metrics of companies. In addition to simple imputation techniques, we also use machine learning techniques that allow working with more complex data.
Methodology- We first employ ad-hoc methods such as mean, median, mode, constant, most frequent and regression imputation. In what follows, we handle multivariate imputation techniques such as multiple imputation by chained equations (MICE). Finally, we run imputation methods with machine learning (ML) classification such as K-nearest Neighbor (KNN), Ridge and Random Forest. To consider the assumptions of missing data, we first check the normality of the variables with Kolmogorov-Smirnov test and employ Rubin’s classification technique that defines the relationship among variables with the probability of missing data. The success of imputation techniques applied to missing data changes when the missing data are classified with Rubin’s technique according to randomness. Consequently, we apply listwise deletion at various levels and alternative data imputation techniques. We then compare their performances. The raw data contain parametric as well as categorical variables (binary and others). Among these are time-series (yearly) financial series such as sales and total assets obtained from financial statements, ESG scores as well as float ratios for firms from several countries and industries. Imputation is done randomly on a sample varying from 5% to 30% of the dataset and results are compared to true data based on accuracy or other measures such as root mean square errors (RMSE) or mean absolute percentage error (MAPE). Several robustness checks have been performed to supplement the analysis.
Findings- Results show that ML methods such as KNN have a superior performance than others. Moreover, when multidimensional nature of the data is taken into account, the prediction performance improves. Hence, an optimality can be reached based on parameters.
Conclusion- Based upon the analysis, we conclude that the selected imputation technique and how it is employed matter to attain a higher accuracy and a better prediction of the missing values on selected mixed panel data in finance.

Keywords: Imputation techniques, Panel data, Machine learning, Financial performance, ESG
JEL Codes: C55, C81, M14, Q51

REFERENCES

Demirtas, H., Freels, S. A., and Yucel, R. M. (2008). Plausibility of multivariate normality assumption when multiply imputing non-Gaussian continuous outcomes: A simulation assessment. Journal of Statistical Computation and Simulation, 78(1), 69–84.

Lin, W. C., & Tsai, C. F. (2020). Missing value imputation: a review and analysis of the literature (2006–2017). Artificial Intelligence Review, 53, 1487–1509.

Little, R. J., & Rubin, D. B. (2020). Statistical Analysis with Missing Data. 3rd ed., John Wiley & Sons.

Rubin, D. B. (1976). Inference and missing data. Biometrika, 63(3), 581–592.

Sahin, Ö., Bax, K., Czado, C., & Paterlini, S. (2022). Environmental, Social, Governance scores and the Missing pillar—Why does missing information matter?. Corporate Social Responsibility and Environmental Management, 29(5), 1782–1798.

Schafer, J. L. (1997). Analysis of Incomplete Multivariate Data. CRC Press.

Serafeim, G. (2015). Integrated reporting and investor clientele. Journal of Applied Corporate Finance, 27(2), 34–51.

Song, Q., & Shepperd, M. (2007). Missing data imputation techniques. International Journal of Business Intelligence and Data Mining, 2(3), 261–291.

Uyar, A., Kuzey, C., & Karaman, A. S. (2022). ESG performance and CSR awards: Does consistency matter?. Finance Research Letters, 50, 103276.

Uyar, A., Kuzey, C., Kilic, M., & Karaman, A. S. (2021). Board structure, financial performance, corporate social responsibility performance, CSR committee, and CEO duality: Disentangling the connection in healthcare. Corporate Social Responsibility and Environmental Management, 28(6), 1730–1748.

Van Buuren, S. (2018). Flexible imputation of missing data. CRC press.

Young, R., & Johnson, D. R. (2015). Handling missing values in longitudinal panel data with multiple imputation. Journal of Marriage and Family, 77(1), 277–294.

Zhang, Z. (2016). Missing data imputation: focusing on single imputation. Annals of Translational Medicine, 4(1):9.