Journal of Quality Measurement and Analysis *JQMA* 21(3) 2025, 1-19
e-ISSN: 2600-8602 *http://www.ukm.my/jqma*
https://doi.org/10.17576/jqma.2103.2025.01

MISSING VALUES TREATMENT IN AGRONOMY DATASET USING PCA-BASED MULTIPLE IMPUTATION (BOOTSTRAP VERSUS BAYESIAN)

(*Rawatan Nilai Lenyap dalam Set Data Agronomi Menggunakan Imputasi Berganda Berasaskan PCA (Bootstap Lawan Bayesian)*)

RAHIMAH SALLEHUDDIN* & NORSHAHIDA SHAADAN

ABSTRACT

Missing values are prevalent in agronomy datasets and need consideration to ensure the applicability of statistical methods and avoid bias in treating them. Previous studies indicate that multiple imputation is more effective than single imputation, with Principal Component Analysis (PCA)-based methods effectively handling multicollinearity in multivariate data. However, such approaches are rarely applied to agronomy data, hence there is a need to assess their performance to add knowledge in the area. This study evaluates the performance of two PCA-based multiple imputation approaches on missing multivariate agronomy data: multiple imputation using regularised PCA through bootstrap procedure (BootMI-REM-PCA) and multiple imputation using regularised PCA through Bayesian procedure (BayesMI-REM-PCA). The data were obtained from the Department of Agriculture Sarawak. A simulation study was conducted using 500 simulated datasets at 5%, 10%, and 20% missingness. Results showed comparable performance between BootMI-REM-PCA and BayesMI-REM-PCA at 5% missingness, with equal coefficient of determination (R^2) values of 0.998, while BootMI-REM-PCA exhibited slightly lower root mean squared error (RMSE) of 1.527 and mean absolute error (MAE) of 0.160. However, BayesMI-REM-PCA outperformed at higher missing rates, achieving the lowest RMSE (2.238 at 10% and 3.051 at 20%) and MAE (0.315 at 10% and 0.601 at 20%), along with the highest R^2 values of 0.996 and 0.993, respectively. While imputation accuracy declines as missing data increases, BayesMI-REM-PCA preserves the characteristics of real data. The findings are expected to help agricultural scientists and researchers prepare high-quality data for accurate analysis.

Keywords: agronomy data, missing data imputation, multiple imputation, PCA-based imputation, multicollinearity

ABSTRAK

Nilai lenyap merupakan isu lazim dalam bidang agronomi yang perlu diambilkira bagi memastikan kepenggunaan kaedah statistik dan mengelakkan bias. Kajian terdahulu menunjukkan kaedah imputasi berganda lebih berkesan berbanding imputasi tunggal, dengan kaedah berasaskan analisis komponen utama (PCA) mampu menangani masalah multikolineariti dalam data multivariat. Namun, kaedah ini jarang diaplikasikan pada data agronomi. Prestasi kaedah ini perlu dinilai untuk menambah pengetahuan dalam bidang ini. Kajian ini menilai prestasi dua kaedah imputasi berganda berasaskan PCA terhadap data agronomi multivariat yang lenyap: imputasi berganda berasaskan PCA menggunakan prosedur bootstrap (BootMI-REM-PCA) dan imputasi berganda berasaskan PCA menggunakan prosedur Bayesian (BayesMI-REM-PCA). Data untuk kajian ini diperoleh daripada Jabatan Pertanian Sarawak. Kajian simulasi dijalankan menggunakan 500 set data yang dijana pada kadar kelenyapan 5%, 10%, dan 20%. Keputusan menunjukkan prestasi setanding antara BootMI-REM-PCA dan BayesMI-REM-PCA pada kadar kelenyapan 5%, dengan nilai pekali penentuan (R^2) yang sama (0.998) dan BootMI-REM-PCA menunjukkan ralat min punca kuasa (RMSE) 1.527 dan ralat min mutlak (MAE) 0.160 yang sedikit lebih rendah. Namun, prestasi BayesMI-

REM-PCA adalah lebih baik pada kadar kelenyapan lebih tinggi dengan RMSE terendah (2.238 pada 10% dan 3.051 pada 20%) dan MAE terendah (0.315 pada 10% dan 0.601 pada 20%), serta nilai R^2 tertinggi (0.996 dan 0.993). Walaupun ketepatan imputasi menurun apabila kadar kelenyapan data meningkat, BayesMI-REM-PCA berkesan mengekalkan ciri-ciri data sebenar. Penemuan ini dijangka dapat membantu penyelidik menyediakan data agronomi yang berkualiti untuk analisis yang tepat.

Kata kunci: data agronomi, imputasi data lenyap, imputasi berganda, imputasi berasaskan PCA, multikolineariti

References

Abbasi M., Yaghmaee M.H. & Rahnama F. 2019. Internet of Things in agriculture: A survey. *Proceedings of the 3rd International Conference on Internet of Things and Applications*, pp. 1-12.

Alwateer M., Atlam E.S., Abd El-Raouf M.M., Ghoneim O.A. & Gad I. 2024. Missing data imputation: A comprehensive review. *Journal of Computer and Communications* **12**(11): 53-75.

Arciniegas-Alarcón S., García-Peña M. & Rodrigues P.C. 2020. New multiple imputation methods for genotype-by-environment data that combine singular value decomposition and Jackknife resampling or weighting schemes. *Computers and Electronics in Agriculture* **176**: 105617.

Arciniegas-Alarcón S., Garcia-Peña M., Krzanowski W.J. & Rengifo C. 2023. Missing value imputation in a data matrix using the regularisuied singular value decomposition. *MethodsX* **11**: 102289.

Audigier V., Husson F. & Josse J. 2016. Multiple imputation for continuous variables using a Bayesian principal component analysis. *Journal of Statistical Computation and Simulation* **86**(11): 2140-2156.

Austin P.C., White I.R., Lee D.S. & van Buuren S. 2021. Missing data in clinical research: A tutorial on multiple imputation. *Canadian Journal of Cardiology* **37**(9): 1322-1331.

Ayilara O.F., Zhang L., Sajobi T.T., Sawatzky R., Bohm E. & Lix L.M. 2019. Impact of missing data on bias and precision when estimating change in patient-reported outcomes from a clinical registry. *Health and Quality of Life Outcomes* **17**: 106.

Ben Aissia M.A., Chebana F. & Ouarda T.B.M.J. 2017. Multivariate missing data in hydrology – Review and applications. *Advances in Water Resources* **110**: 299-309.

Bertsimas D., Pawlowski C. & Zhuo Y.D. 2018. From predictive methods to missing data imputation: An optimisation approach. *Journal of Machine Learning Research* **18**(196): 1-39.

Brini A. & van den Heuvel E.R. 2024. Missing data imputation with high-dimensional data. *The American Statistician* **78**(2): 240–252.

Chaudhry A., Li W., Basri A. & Patenaude F. 2019. A method for improving imputation and prediction accuracy of highly seasonal univariate data with large periods of missingness. *Wireless Communications and Mobile Computing* **2019**(1): 4039758.

Chicco D., Warrens M.J. & Jurman G. 2021. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Comput. Sci.* **7**: e623.

Chow C., Andrášik R., Fischer B. & Keiler M. 2019. Application of statistical techniques to proportional loss data: Evaluating the predictive accuracy of physical vulnerability to hazardous hydro-meteorological events. *Journal of Environmental Management* **246**: 85-100.

Dubey A. & Rasool A. 2020. Clustering-based hybrid approach for multivariate missing data imputation. *International Journal of Advanced Computer Science and Applications* **11**(11): 710-714.

Faisal S. & Tutz G. 2021. Multiple imputation using nearest neighbour methods. *Information Sciences* **570**: 500-516.

Fang L., Xiang W., Zhou Y., Fang J., Chi L. & Ge Z. 2023. Dual-branch cross-dimensional self-attention-based imputation model for multivariate time series. *Knowledge-Based Systems* **279**: 110896.

Fioroni N., Mouquet-Rivier C., Meudec E., Cheynier V., Boudard F., Hemery Y. & Laurent-Babot C. 2023. Antioxidant capacity of polar and non-polar extracts of four African green leafy vegetables and correlation with polyphenol and carotenoid contents. *Antioxidants* **12**(9): 1726.

Fu Y., Liao H. & Lv L. 2021. A comparative study of various methods of handling missing data in UNSODA. *Agriculture* **11**(8): 727.

Gad A., Malouche D., Chhabra M., Hoang D., Suk D., Ron N., Dygulska B., Gudavalli M.B., Nadroo A.M., Narula P. & Elmakaty I. 2024. Impact of birth weight to placental weight ratio and other perinatal risk factors on left ventricular dimensions in newborns: A prospective cohort analysis. *Journal of Perinatal Medicine* **52**(4): 433-444.

Gomez-Bougie P., Maiga S., Tessoulin B., Bourcier J., Bonnet A., Rodriguez M.S., Le Gouill S., Touzeau C., Moreau P., Pellat-Deceunynck C. & Amiot M. 2018. BH3-mimetic toolkit guides the respective use of BCL2 and MCL1 BH3-mimetics in myeloma treatment. *Blood, The Journal of the American Society of Hematology* **132**(25): 2656-2669.

Graham J.W., Olchowski A.E. & Gilreath T.D. 2007. How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention science* **8**: 206-213.

Gwelo A.S. 2019. Principal components to overcome multicollinearity problem. *Oradea Journal of Business and Economics* **4**(1): 79-91.

Hughes R.A., Heron J., Sterne J.A.C. & Tilling K. 2019. Accounting for missing data in statistical analyses: Multiple imputation is not always the answer. *International Journal of Epidemiology* **48**(4): 1294-1304.

Jinubala V. & Lawrance R. 2016. Analysis of missing data and imputation on agriculture data with predictive mean matching method. *International Journal of Science and Applied Information Technology* **5**(1): 01-04.

Jollife I.T. & Cadima J. 2016. Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **374**(2065): 20150202.

Jose J., Vishwakarma G.K. & Bhattacharjee A. 2021. Illustration of missing data handling technique generated from hepatitis C induced hepatocellular carcinoma cohort study. *Journal of King Saud University - Science* **33**(4): 101403.

Josse J. & Husson F. 2012. Handling missing values in exploratory multivariate data analysis methods. *Journal de la Société Française de Statistique* **153**(2): 79-99.

Josse J. & Husson F. 2016. missMDA: A package for handling missing values in multivariate data analysis. *Journal of Statistical Software* **70**(1): 1-31.

Josse J., Pagès J. & Husson F. 2011. Multiple imputation in principal component analysis. *Advances in Data Analysis and Classification* **5**: 231-246.

Kim J.H. 2019. Multicollinearity and misleading statistical results. *Korean Journal of Anesthesiology* **72**(6): 558–569.

Kim W., Cho W., Choi J., Kim J., Park C. & Choo J. 2019. A comparison of the effects of data imputation methods on model performance. *Proceedings of the 21st International Conference on Advanced Communication Technology (ICACT)*, pp. 592-599.

Li P. & Stuart E. A. 2019. Best (but oft-forgotten) practices: Missing data methods in randomised controlled nutrition trials. *The American Journal of Clinical Nutrition* **109**(3): 504-508.

Liu G., Yang Y. & Li B. 2018. Fuzzy rule-based oversampling technique for imbalanced and incomplete data learning. *Knowledge-Based Systems* **158**: 154-174.

Lohr S.L. 2010. *Sampling: Design and Analysis*. 2nd Ed. Belmont, CA: Brooks/Cole.

Loisel S. & Takane Y. 2019. Comparisons among several methods for handling missing data in principal component analysis (PCA). *Advances in Data Analysis and Classification* **13**: 495-518.

Maliwal P.L. & Mundra S.L. 2011. *Agronomy at a Glance: Vol-2 Objective Fundamentals*. 3rd Ed. Udaipur, Rajasthan, India: Agrotech Publishing Academy.

Mensching A., Hummel J. & Sharifi A.R. 2020. Statistical modeling of ruminal pH parameters from dairy cows based on a meta-analysis. *Journal of Dairy Science* **103**(1): 750-767.

Mensching A., Zschiesche M., Hummel J., Grelet C., Gengler N., Dänicke S. & Sharifi A.R. 2021. Development of a subacute ruminal acidosis risk score and its prediction using milk mid-infrared spectra in early-lactation cows. *Journal of Dairy Science* **104**(4): 4615-4634.

Noghrehchi F., Stoklosa J., Penev S. & Warton D.I. 2021. Selecting the model for multiple imputation of missing data: Just use an IC!. *Statistics in Medicine* **40**(10): 2467-2497.

Roney P.R., Lai L.S., Hamsein N.N. & Sallehudin R. 2023. Broccoli (Brassica oleraceae var. italica) growth performance in lowland using fertigation under protected rain shelter: The influence of spacings and accessions. *Malaysian Applied Biology* **52**(5): 205-211.

Royston P. & White I.R. 2011. Multiple imputation by chained equations (MICE): Implementation in Stata. *Journal of statistical software* **45**(4): 1-20.

Saeipourdizaj P., Sarbakhsh P. & Gholampour A. 2021. Application of imputation methods for missing values of PM10 and O3 data: Interpolation, moving average and K-nearest neighbour methods. *Environmental Health Engineering and Management Journal* **8**(3): 215-226.

Sanju, Kumar V. & Deepender. 2023. Evaluation of imputation techniques for genotypic data of soybean crop under missing completely at random mechanism. *Indian Journal of Agricultural Research* **57**(5): 701-705.

Soley-Bori M. 2013. Dealing with missing data: Key assumptions and methods for applied analysis. Technical Report. School of Public Health, Boston University.

Solfanelli F., Gambelli D., Vairo D. & Zanoli R. 2019. Estimating missing data for organic farming by multiple imputation: The case of organic fruit yields in Italy. *Organic Agriculture* **9**: 295-303.

Stochero E.L.M., Lúcio A.D.C. & Jacobi L.F. 2024. Data variability in the imputation quality of missing data. *Acta Scientiarum. Agronomy* **46**: e66185.

Waljee A.K., Mukherjee A., Singal A.G., Zhang Y., Warren J., Balis U., Marrero J., Zhu J. & Higgins P.D. 2013. Comparison of imputation methods for missing laboratory data in medicine. *BMJ Open* **3**(8): e002847.

Wani A.A. 2024. A review of challenges and solutions for using machine learning approaches for missing data. *International Journal of Engineering Applied Sciences and Technology* **9**(5): 36-50.

White I.R., Royston P. & Wood A.M. 2011. Multiple imputation using chained equations: issues and guidance for practice. *Statistics in Medicine* **30**(4): 377-399.

Wiangsamut B. & Koolpluksee M. 2020. Yield and growth of Pak Choi and Green Oak vegetables grown in substrate plots and hydroponic systems with different plant spacing. *International Journal of Agricultural Technology* **16**(4): 1063-1076.

Wimalasena S.T., Ramirez Silva C.I., Sun Y.V., Stein A.D., Rivera J.A. & Ramakrishnan U. 2023. Clustering of cardiometabolic risk factors in Mexican pre-adolescents. *Diabetes Research and Clinical Practice* **202**: 110818.

You J., Ellis J.L., Adams S., Sahar M., Jacobs M. & Tulpan D. 2023. Comparison of imputation methods for missing production data of dairy cattle. *animal* **17**: 100921.

Zhong H., Hu W. & Penn J.M. 2018. Application of multiple imputation in dealing with missing data in agricultural surveys: The case of BMP adoption. *Journal of Agricultural and Resource Economics* **43**(1): 78-102.

Agriculture Research Centre Semongok
Department of Agriculture Sarawak
P.O.Box 977, 93720 Kuching,
Sarawak, MALAYSIA
E-mail: rahimahs@sarawak.gov.my[*]

College of Computing, Informatic, and Mathematics
University Technology MARA
40450 Shah Alam,
Selangor, MALAYSIA
E-mail: shahida@tmsk.uitm.edu.my

Received: 16 December 2024
Accepted: 4 April 2025

[*]Corresponding author