

## DISTRIBUTION-FREE MULTIPLE IMPUTATION IN AN INTERACTION MATRIX, USING THE SINGULAR VALUE DECOMPOSITION

Carlos Tadeu dos Santos Dias<sup>1</sup>, Genevile Carife Bergamo<sup>2</sup> \*; Wojtek Janusz Krzanowski<sup>3</sup>

<sup>1</sup>Departamento de Ciências Exatas, ESALQ, Universidade de São Paulo - USP, Caixa Postal 09, CEP: 13418-900, Piracicaba, SP, Brasil.

<sup>2</sup>Universidade José do Rosário Velano - UNIFENAS, Caixa Postal 23, CEP: 37130-000, Alfenas, MG, Brasil.

<sup>3</sup>School of Engineering, Computer Science & Mathematics, University of Exeter, North Park Road, Exeter, EX4 4QE, UK.

**ABSTRACT:** Some techniques of multivariate statistical analysis can only be conducted on a complete data matrix, but the process of data collection is often such that some elements are missing. Imputation is a technique in which the missing elements are replaced by plausible values, so that a valid analysis can be done on the completed data set (observed + imputed). We here propose a multiple imputation method based on a modification to the singular value decomposition (SVD) method for single imputation developed by Krzanowski (1988), and evaluate it on the genotype  $\times$  environment ( $G \times E$ ) interaction matrix obtained in a randomized blocks experiment on the *Eucalyptus grandis* genotype in multienvironments (Lavoranti, 2003). Values of *E. grandis* heights in the  $G \times E$  interaction matrix were deleted randomly at three different rates (5%, 10%, 30%) and were then imputed by the proposed method. The results were assessed by means of a general measure of performance ( $T_{acc}$ ), and showed a small bias when compared to the original data. However, bias values were greater than the variability of imputations relative to their mean, indicating a smaller accuracy of the proposed method in relation to its precision. The proposed methodology uses the maximum amount of information available, it does not have any restrictions regarding the pattern or mechanism of the missing values, and it is free of assumptions about the data distribution or structure.

**Key words:** Missing data, Nonparametric, Eigenvalue, Eigenvector, Genotype-environment