Comparative study of single imputation techniques for prediction of missing dairy data

Abstract:

Dairy Farm Record is a necessary element of good livestock business management. record analysis allows a farm owner to make informed decisions based on complete records. Because incomplete records are less valuable for data analysis, it is critical to deal with missing value issue appropriately. The current study aimed to compare imputation methods for the estimation of missing values in a raw dataset of dairy cattle including 997 records collected from 234 cows between 2012 and 2022. The dataset was screened against records with missing values then deleted, reducing its size to 858 observations equivalent to 200 cows. There were missing values in two variables with a missing percentage 13.9%: days in milk (DIM) and total milk yield (TOTM). Then, excluding observations with known values at random that exhibit the same missing data percentages as the original dataset for DIM and TOTM. Five different imputation methods were compared to obtaining the best imputation technique for prediction of missing values. These methods are mean imputation, median imputation, power regression imputation, multiple regression imputation and expectation maximization method (EM). The five methods were evaluated based on four performance metrics: mean absolute deviation (MAD), mean square error (MSE), Spearman's rank correlation coefficient (r_s) and mean absolute percentage error (MAPE). The results showed that the expectation maximization method was overall the best imputation method for data under study and had the lowest MAD, the lowest MSE, the highest Spearman's correlation coefficient and the second lowest MAPE for predicting missing dairy cow DIM and TOTM data.

Keywords: record keeping, missing data, imputation methods, expectation maximization, power regression.