EGU26-1739
EGU General Assembly 2026

Missing data imputation in epidemiology: a comparison between MICE and Machine Learning methods

Mahmoud Hashoush[1], Emmanuelle Cadot[1], and Franco Alberto Cardillo[2]
[1]Univ Montpellier, CNRS, IRD, Montpellier, France
[2]Istituto di Linguistica Computazionale, Consiglio Nazionale delle Ricerche, Italia

Missing data represents a challenge in large-scale epidemiological studies as it can introduce a strong and negative bias in the final estimates when not handled appropriately. This issue is particularly relevant in environment health research due to complex relationships between the exposure to risk factors and delayed outcomes. In this work, we evaluate the effectiveness of statistical and Machine Learning (ML) approaches to fill in missing values in data we collected to assess the potential impact on public health of gold mining activities in the Ecuadorian Amazon.

There is growing concern regarding the adverse effects on human health in the Ecuadorian Amazon caused by the environmental impact of gold mining activities in the area. To investigate potential associations with adverse birth outcomes, we collected data published by the Ecuadorian National Institute of Statistics and Census (INEC) relative to the annual live birth and fetal death cases in the years from 2014 to 2023. As it is typical in large-scale epidemiological studies, the data contain a proportion of missing values, likely related to the registration and the data entry process.

Addressing missing values is considered important for the correct assignment of cases from one hand and the characterisation of risk factors from another. Furthermore, it enables the modelling process when searching for associations between exposure and outcome without erroneous under- or over-reporting of odds ratios (Type I and Type II errors). Currently, the most common approach in epidemiology is to use statistical methods and, specifically, Multivariate Imputation by Chained Equations (MICE), normally instantiated with parametric conditional models. MICE imputes missing values by repeatedly predicting each incomplete variable from the others using standard regression models. In most applications, these predictions rely on linear or generalised linear relationships between variables. This can reduce its effectiveness in predicting missing values in presence of complex, non-linear interactions about variables. Machine Learning represents an interesting alternative as it capture complex, non-linear relationships beyond the linear models typically assumed in MICE, are more flexible with respect to departures from missing-at-random patterns, and reduce the risk of model misspecification by relying on data-driven, implicit model selection rather than requiring the analyst to pre-specify an imputation model.

In this study, we present a robust experimental comparison between MICE and several ML-based imputation approaches applied to the Ecuadorian birth data. We assess their performance and

discuss the respective strengths and limitations within an epidemiological context.