Stellenbosch UNIVERSITY IYUNIVESITHI UNIVERSITEIT

Statistics and Actuarial Science
Statistiek en Aktuariële Wetenskap

NITheCS
National Institute for Theoretical and Computational Sciences

S E M I N A R

James Grove
(Stellenbosch University)

James is a Research and Development Analyst at Dynamo, based in Stellenbosch. James holds a BCom in Actuarial Science, an Honours degree in Mathematical Statistics, and a Master's in Statistics and Data Science.

Tree-based imputations in the presence of right-censored data for machine learning applications

ABSTRACT:

Machine learning has many applications, but relatively few methods have been developed for scenarios where censoring occurs in the response. The most straightforward strategy, complete-data analysis, involves restricting model fitting to observations with fully observed responses. While simple, this approach often introduces substantial bias, and discards partial information contained in censored observations.

Imputation is a valuable alternative method for handling censored data. It is a model-agnostic approach in which values are generated to replace censored observations while ensuring that each imputed value exceeds its corresponding censoring point. Traditional imputation techniques often rely on the marginal distribution of the response and do not incorporate covariate information, limiting their usefulness in a predictive modelling context. To address this gap, recursively imputed survival trees were proposed in the literature, which use extremely randomised survival trees to estimate the conditional distribution of the response given the covariates. Imputations are then drawn from this conditional distribution, truncated at the censoring point. In this thesis, we propose a related methodology. The terminal nodes of tree-based methods were used to obtain an estimate of the distribution of the response given the covariates. Imputations are then performed using these estimated distributions. Because the models used for imputation are trained by minimising the squared error of predicted values, they may yield more accurate imputations compared to approaches based on survival functions. Additionally, the widespread familiarity and interpretability of standard regression trees may enhance the accessibility and usability of the proposed methodology. Although the methodology can be applied for general censoring, the focus of this thesis, is for right-censoring.

We compared our proposed method to complete-data analysis, adjusting the sampling weights with the inverse probability of censoring, random survival forests, and a benchmark model trained on fully observed data (possible in controlled settings). They were compared in a simulation study as well as with real-world data in which soil depths in SA were modelled. The results demonstrated that the proposed method performs competitively across different censoring scenarios.

Date:
Friday, 13 March 2026

Time:
13h10-14h10 SAST

Venues:
- Room 2048
  Van der Sterr Building,
  cnr Victoria & Bosman Streets
  Stellenbosch
- Online

WHO SHOULD ATTEND?
All are welcome.

ENQUIRIES:

Elizna Huysamen
☎ +27 (0)21 808 3244
✉ krugere@sun.ac.za

REGISTER: https://bit.ly/3ZYUJwV

science, technology & innovation
Department:
Science, Technology and Innovation
REPUBLIC OF SOUTH AFRICA

NRF National Research Foundation