Applied Data Science (ADS) project acquisition form

Title of the project *

Can the best prediction be the worst imputation?

Number of students for the project: (typically, projects have 2-3 students) *

3

Description (abstract size, approximately 200 words) *

Missing values often complicate the analysis of data. A technique that solves this problem is imputation. With imputation, we first solve the problem of missingness by filling in values for the unobserved cells. We then analyze the data as if the data were completely observed. If we would do this correct in a statistical sense, we need to fill in multiple values for each missing datum to reflect the uncertainty associated with the missingness problem. From a mathematical point of view, filling in the best single prediction would be most efficient. In a prediction problem where both the outcome and the predictors have unobserved values, both approaches may yield different solutions when answering the problem formulated on the data. To what extend these approaches differ is aimed to be investigated in the following questions:

1. If the goal is prediction, what would be the best imputation out of
    a. Mean imputation (fastest)
    b. Regression imputation (fast)
    c. Stochastic regression imputation (slow)
    d. Parametric imputation (slow)
    e. Non-parametric imputation (slowest)
2. Would – for the above techniques c, d and e - there be any benefit to perform the imputations multiply?
3. Does a higher level of missingness yield a different preferred imputation technique?

Literature is available. No previous experience with incomplete data theory is required, students will receive a small private self-paced course in incomplete data theory to get them up to speed.

Organization name and names of internal supervisors involved. *

Utrecht University / Social Sciences / Department of Methodology and Statistics /

Names of supervisors from Utrecht University

Stef van Buuren / Gerko Vink / Hanne Oberman

Website address for additional information of organization or project *

https://github.com/amices

https://stefvanbuuren.name/fimd/

Short description of the available data. *

Data generated by simulating missing values on a prepared data set for which the true data generating model is known.

Project domain *

Social and behavioural science

Optional: required courses in domain https://www.uu.nl/masters/en/applied-data-science/courses

Epidemiology and Big Data
Using data from routine care, registries, health devices and public repositories
Spatial data analysis and simulation modelling
Spatial Statistics and Machine Learning
Social Behavioural Dynamics
Network Analysis
Data Mining: Text, Images, Video
Personalisation for (Public) Media

Additional requirements (such as signing an NDA, clearance, etc.)

None

Optional: Add a pdf/word document with extra details