Applied Data Science (ADS) project acquisition form

Title of the project *

Modeling not applicable answers when data are incomplete.

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

3

Description (abstract size, approximately 200 words) *

Many survey efforts allow for a distinct answer category that models when a question or answer would be `not applicable`. A `not applicable` is a bonafide answer category, even though this may seem counterintuitive in practice. As a consequence, modeling the observed data becomes increasingly challenging when not applicable's are encountered, which is further complicated when not all data are observed. How should anwers be found to questions on incomplete data that exhibits not applicable anwer categories? To investigate this, we answer the following questions in simulations:

1. What happens to the validity of inferences if we simply ignore the observed not applicables?
2. What happens to the validity of inferences if we simply ignore the unobserved not applicables?
3. Is it better to model unobserved not applicables in a single step (e.g. with non-parametric techniques), or should we use parametric techniques two-step techniques that first model whether or not a value should be not-applicable or otherwise?
4. Does stratified analysis of not-applicable patterns provide a useful answer to the data problem?

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