Contribution ID: **19** Type: **not specified**

Poster: Sampling to adjust for imbalance in production data.

Production often has as main purpose that products should fall in a pre-specified range of variation. Thus, there is a lack of variation in data making it difficult to work with from a statistical viewpoint. The traditional approach to learn about the input-to-output process of a production is to make a hypothesis and then tailor an experimental design yielding sufficient variation in data to test the hypothesis. However, this is costly and sometimes impossible. Instead we have to settle with the historical data.

Due to the purpose of production, most data lie in a high-density area, with only a few points falling outside. This is denoted imbalance and is only sparsely studied in continuous data, whereas the discrete counterpart has been studied extensively. Only imbalance in continuous data in regard to the output has previously been addressed.

We study imbalance in regard to the covariate domain. The study is a simulation study. We evaluate different approaches of pre-processing data aiming at a more balanced data set. It is found that in lower-to-medium dimensional settings, fitting the model with the pre-processed data yields a better predictive performance than using the entire data to fit the model.

Primary author: GRØNBERG, Manja (Technical University of Denmark)

Co-authors: SVENDSEN, Kira Dynnes; MÅGE, Ingrid; CLEMMENSEN, Line Katrine Harder (Technical University of Denmark)

Presenter: GRØNBERG, Manja (Technical University of Denmark)

Track Classification: Data Science in Process Industries