Engineering Risk Analysis Group
Prof. Dr Daniel Straub

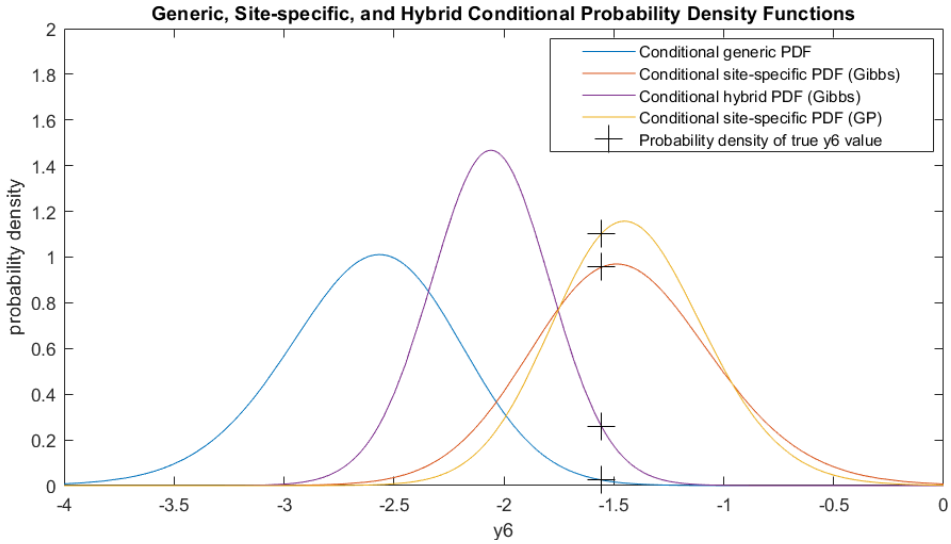
MSc thesis

Construction and Evaluation of Machine Learning Models for Predicting a Multivariate Probability Distribution of Design Soil Parameters

Dominik Hesping, January 2020

Background

Measuring subsurface soil parameters is an essential step in the investigation process of construction sites. The research of probabilistic models to estimate important design soil parameters, like the undrained shear strength (s_u), has increased due to the availability of large geological databases. Especially the distinguished professors Jianye Ching and Kok-Kwang Phoon did some major work in this field and the thesis is based on their models. The novel part of the thesis is the evaluation of the predictive performance of those models for estimating s_u, and the comparison with other machine learning models constructed in this thesis.



Conditional Probability Density Functions (PDFs) of the undrained shear strength (s_u) given a set of measured predictor parameters. The PDFs are constructed using a generic, a Gibbs sampling, a hybrid, and a Gaussian processes approach and the true s_u values are marked

Methodology

Generic models which use global soil data are constructed using different approaches for estimating the marginal distributions (e.g. Johnson distributions or kernel density estimation) and a Gaussian copula. Site-specific models purely rely on the data of single sites and their statistical uncertainty needs to be reflected. A Gibbs sampling and a Gaussian processes model are constructed. The evaluation of the predictive performance is done with cross-validation and the error measure is chosen to be the mean cross-entropy loss of all measurements.

Conclusion

For the generic models, using all ten soil parameters of the highly incomplete data set leads to overconfident predictions and non-positive definite matrices. A possible solution is shown using small perturbations of the uncertain correlations. Generic models perform best for sites with sparse data. The Gibbs sampling method performs best in general and is especially suited for abundant data. The Gaussian processes approach needs to be improved and performs well for some sites with a medium amount of measurements.

Supervised by M.Sc. Marco Kryda, Dr.-Ing. Iason Papaioannou, Prof. Dr. Daniel Straub