June 7-12, 2026
Nanjing, China
https://www.23wcss.org.cn/

Evaluating Conditioned Probabilistic Subsampling for the Validation of 3D Regional Subsurface Models using Legacy Data

Pablo De Weerdt[1*], Stijn Luca[2], and Ellen Van De Vijver[1]

[1] *Department of Environment, Ghent University, Ghent, Belgium*

[2] *Department of Data Analysis and Mathematical Modelling, Ghent University, Ghent, Belgium*

*Pablo.DeWeerdt@Ugent.be

ABSTRACT

With the increasing availability of soil data in various forms it is more common to develop models using only legacy data. While model development is ideally preceded by sampling design aligned with the desired outcome, pre-existing data can partially fulfil requirements and allow for larger-scale modelling such as regional subsurface characterisation. The resulting models can also serve as tools to identify information gaps and model limitations. Interpreted soil borehole profiles, for example, form the basis of multidimensional subsurface models. The availability of data in a certain study area, however, can still be relatively low given the desired resolution or precision of the model. A deliberate decision on sampling existing datasets in general, and creating subsamples for model calibration and validation in particular, is therefore essential during performance assessment of different models. While cross-validation techniques are commonly applied to maximize data utility, their application in spatial modelling yields overoptimistic performance estimates with high variance, particularly when data are clustered. Probabilistic-based sampling is known to tackle bias, but its effectiveness remains poorly understood for spatially sparse and clustered legacy data.

In this research, the effect of subsampling methods on the validation of spatial interpolation techniques is assessed in a study region with legacy boreholes spanning over a century. Simple random subsampling is compared to conditioned Latin hypercube subsampling of boreholes in terms of true model performance. The conditioning 2D legacy characteristics consisted of age, coordinates and target feature statistics. We evaluated the impact of subsample size and anisotropic autocorrelation structure using the same 2D legacy configuration populated with three different 3D target features: two spatially continuous synthetic distributions and one heterogeneous field-observed peat content distribution. Variance reduction was observed only in the spatially heterogeneous case when the validation subsample size was sufficient (35%) to support conditioning that included the cumulative peat content within a borehole. This demonstrates that conditioned probabilistic subsampling can be more robust than other methods for the validation of models derived from legacy data. Our findings also provide insights into different sources of variance within validation results.