Geophysical Research Abstracts
Vol. 19, EGU2017-2175, 2017
EGU General Assembly 2017

Subsampling for dataset optimisation

Mareike Ließ
Helmholtz Centre for Environmental Research - UFZ, Halle (Saale), Germany (mareike.liess@ufz.de)

Soil-landscapes have formed by the interaction of soil-forming factors and pedogenic processes. In modelling these landscapes in their pedodiversity and the underlying processes, a representative unbiased dataset is required. This concerns model input as well as output data. However, very often big datasets are available which are highly heterogeneous and were gathered for various purposes, but not to model a particular process or data space. As a first step, the overall data space and/or landscape section to be modelled needs to be identified including considerations regarding scale and resolution. Then the available dataset needs to be optimised via subsampling to well represent this n-dimensional data space. A couple of well-known sampling designs may be adapted to suit this purpose. The overall approach follows three main strategies: (1) the data space may be condensed and de-correlated by a factor analysis to facilitate the subsampling process. (2) Different methods of pattern recognition serve to structure the n-dimensional data space to be modelled into units which then form the basis for the optimisation of an existing dataset through a sensible selection of samples. Along the way, data units for which there is currently insufficient soil data available may be identified. And (3) random samples from the n-dimensional data space may be replaced by similar samples from the available dataset. While being a presupposition to develop data-driven statistical models, this approach may also help to develop universal process models and identify limitations in existing models.