EGU2020-5249
https://doi.org/10.5194/egusphere-egu2020-5249
EGU General Assembly 2020

Missing data imputation for multisite rainfall networks: a comparison between geostatistical interpolation and data-mining estimation on different terrain types

Fabio Oriani[1], Simon Stisen[2], Mehmet C. Demirel[3], and Gregoire Mariethoz[1]
[1]Faculty of Geosciences and Environment, Institute of Earth Surface Dynamics, University of Lausanne, Switzerland
[2]Department of Hydrology, Geological Survey of Denmark and Greenland, Copenhagen, Denmark
[3]Department of Civil Engineering, Istanbul Technical University, Instanbul, Turkey

In the era of big data, missing data imputation remains a delicate topic for both the analysis of natural processes and to provide input data for physical models. We propose here a comparative study for missing data imputation on daily rainfall, a variable that can exhibit a complex structure composed of a dry/wet pattern and anisotropic sharp variations.

The seven algorithms considered can be grouped in two families: geostatistical interpolation techniques based on inverse-distance weighting and Kriging, widely used in gap-filling [1], and data-driven techniques based on the analysis of historical data patterns. This latter family of algorithms has been already applied to rainfall generation [2, 3], but it is not originally suitable to historical datasets presenting many data gaps. This happens because they usually operate in a rigid framework where, when a rainfall value is estimated for a station, the others are considered as predictor variables and require to be informed. To overcome this limitation, we propose here i) an adaptation of k-nearest neighbor (KNN) and ii) a new algorithm called Vector Sampling (VS), that combines concepts of multiple-point statistics and resampling. These data-driven algorithms can draw estimations from largely and variably incomplete data patterns, allowing the target dataset to be at the same time the training dataset.

Tested on different case studies from Denmark, Australia, and Switzerland, the algorithms show a different performance that seems to be related to the terrain type: on flat terrains with spatially uniform rain events, geostatistical interpolation tends to minimize the error, while, in mountainous regions with non-stationary rainfall statistics, data mining can recover better the complex rainfall patterns. The VS algorithm, being faster than KNN and requiring minimal parametrization, turns out to be a convenient option for routine application if a representative historical dataset is available. VS is open-source and freely available at .

REFERENCES:

org/

org/