Geophysical Research Abstracts
Vol. 18, EGU2016-2306-1, 2016
EGU General Assembly 2016


# Structogram: A new approach to characterize and interpolate spatio-temporal data sets based on the Ramer-Douglas-Peucker algorithm

Uwe Ehret

Karlsruhe Institute of Technology KIT, Institute of Water Resources and River Basin Management, Karlsruhe, Germany (uwe.ehret@kit.edu)

The Ramer-Douglas-Peucker algorithm (Ramer, 1972; Douglas and Peucker, 1973) is a procedure to approximate a polygon of arbitrary dimension (basically any spatial or temporal data set) by a subset of its nodes. Developed in the field of image processing, it is mainly used to compress images while preserving their main features. The core algorithm is straightforward: Starting by approximating the original polygon by a straight line/plane connecting the outside edge nodes, more nodes are successively added in the order of their informativeness about the original field (always the node farthest from the previous field approximation is added) until a desired agreement, expressed by some distance measure such as the mean absolute error, between approximation and original polygon is achieved.

This yields a list of nodes ordered by their relevance, which can be used in various ways to characterize and interpolate data sets. The scope of this talk is to present and discuss several of these ways and to compare them to established (geo-)statistical methods such as Variance, Variogram, and Kriging.

Characterization of data sets

Plotting the number of nodes used against the field approximation error yields what can be called a 'structogram', as it reflects how information on the original data set is distributed among the nodes, or in other words how 'structured' the data set is: In highly structured data sets such as discharge time series, few nodes suffice to represent the original time series very accurately, and adding more nodes does not yield much more improvement, while for unstructured data sets such as white noise fields, each new node reduces the approximation error by a comparable increment. With the structogram, it is also possible to determine a Pareto optimum between the number of nodes used and the corresponding approximation error. For a highly structured data set such as discharge, the Pareto optimum is reached with much less points and has a much lower approximation error than for an unstructured data set such as white noise. Knowledge of this Pareto optimum can be useful for the design of sampling strategies.

It is also interesting to analyze the spatio-temporal distribution of the most relevant nodes of the data set (those with the largest information gain): Homogeneously spaced nodes indicate a data set of constant predictability throughout its extent, or low complexity, while heterogeneously spaced nodes indicate shifting patterns of local predictability, which is an attribute of higher complex data sets (if 'complexity' is defined as 'high overall uncertainty about local uncertainty').

Interpolation of data sets

The structogram can also be used for interpolation, i.e. estimation at nodes where no observations are available. The idea of structogram-based interpolation is that, just as for Kriging, the estimation is a weighted linear combination of the observations, but here the weights are not determined based on the Variogram and the intrinsic hypothesis, but on the relevance of the nodes: Highly relevant nodes are given higher weights than lesser relevant nodes. Testing many different data sets revealed that for 'smooth' data sets, where proximity means similarity, classical Kriging-based interpolation outperforms structogram-based approaches, while for intermittent data sets such as rainfall time-series, where proximity does not always mean similarity, structogram-based interpolation performs better.

References

Ramer, U.: An iterative procedure for the polygonal approximation of plane curves, Computer Graphics and Image Processing, 1, 244-256, http://dx.doi.org/10.1016/S0146-664X(72)80017-0, 1972.

Douglas, D., Peucker, T.: Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. In: The Canadian Cartographer. Bd. 10, Nr. 2, 1973, ISSN 0008-3127, S. 112–122, 1973.