Karlsruhe Institute of Technology

Institute for Water and River Basin Management
Chair of Hydrology

Structogram: A method to quantify the structuredness and complexity of data sets

Uwe Ehret

1 Introduction

This poster presents a method to quantify the structuredness and complexity of data sets based on a Structogram. 'Structuredness' relates to how information on the data set is distributed among its elements, 'complexity' relates to how this information is distributed over the extent of the data set.

2 The Structogram method

The core procedure is to approximate the data set by a subset of its elements and, via (here: linear) interpolation, to estimate the values of all elements Fig. 1).

The related estimation error quantifies the information about the data set contained in the subset. Increasing the subset size from zero by always adding the next most informative element yields an ordered list of subset size and related error, which is the Structogram (Fig. 2). For each data set, two characteristic subsets can be identified: The minimum support subset, beyond which estimates are better than by simply using the mean, and the pareto optimal subset which jointly minimizes subset size and the related error and indicates the optimal level of (lossy) compression for the data set (see red and green dot in Fig. 2). With the pareto set, the 'structuredness' can be calculated:

$$Structuredness = 1 - \frac{points\ in\ pareto\ subset}{points\ in\ data\ set}$$

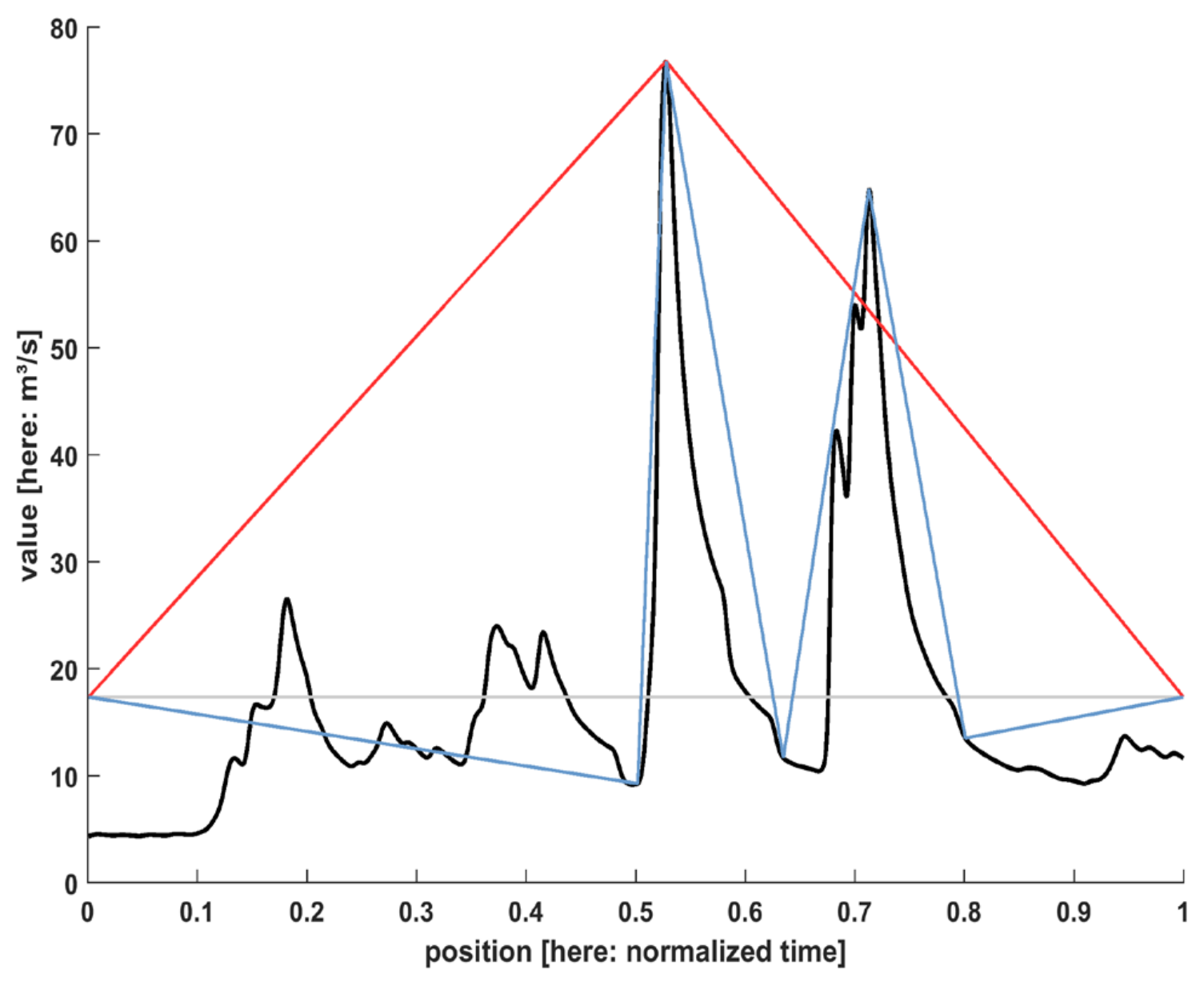

Figure 1: A hydrograph approximated by two (grey), three (red) and seven (blue) points.

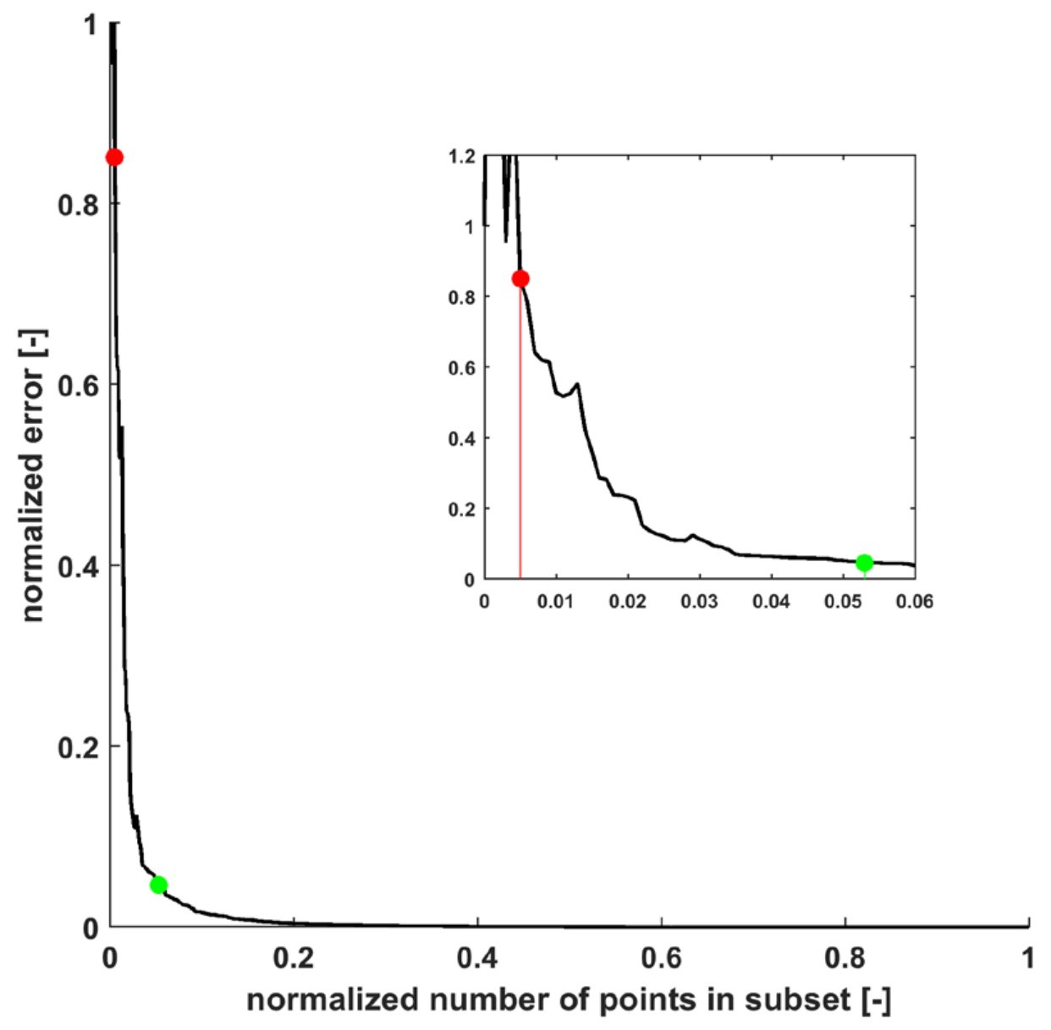

Figure 2: Structogram of the hydrograph in Fig. 1. Red dot = minimum support, green dot = pareto optimum

The Structogram method (cont'd)

The complexity of the data set is estimated by applying the Structogram procedure to the distances between (not the values of) the elements of the pareto subset (Fig. 3). Irregularly spaced elements indicating complex data sets lead to large pareto optimal subsets, little complex data sets lead to small pareto subsets. Complexity is calculated by:

$$Complexity = \frac{points\ in\ distance\ pareto\ subset}{points\ in\ data\ set}$$

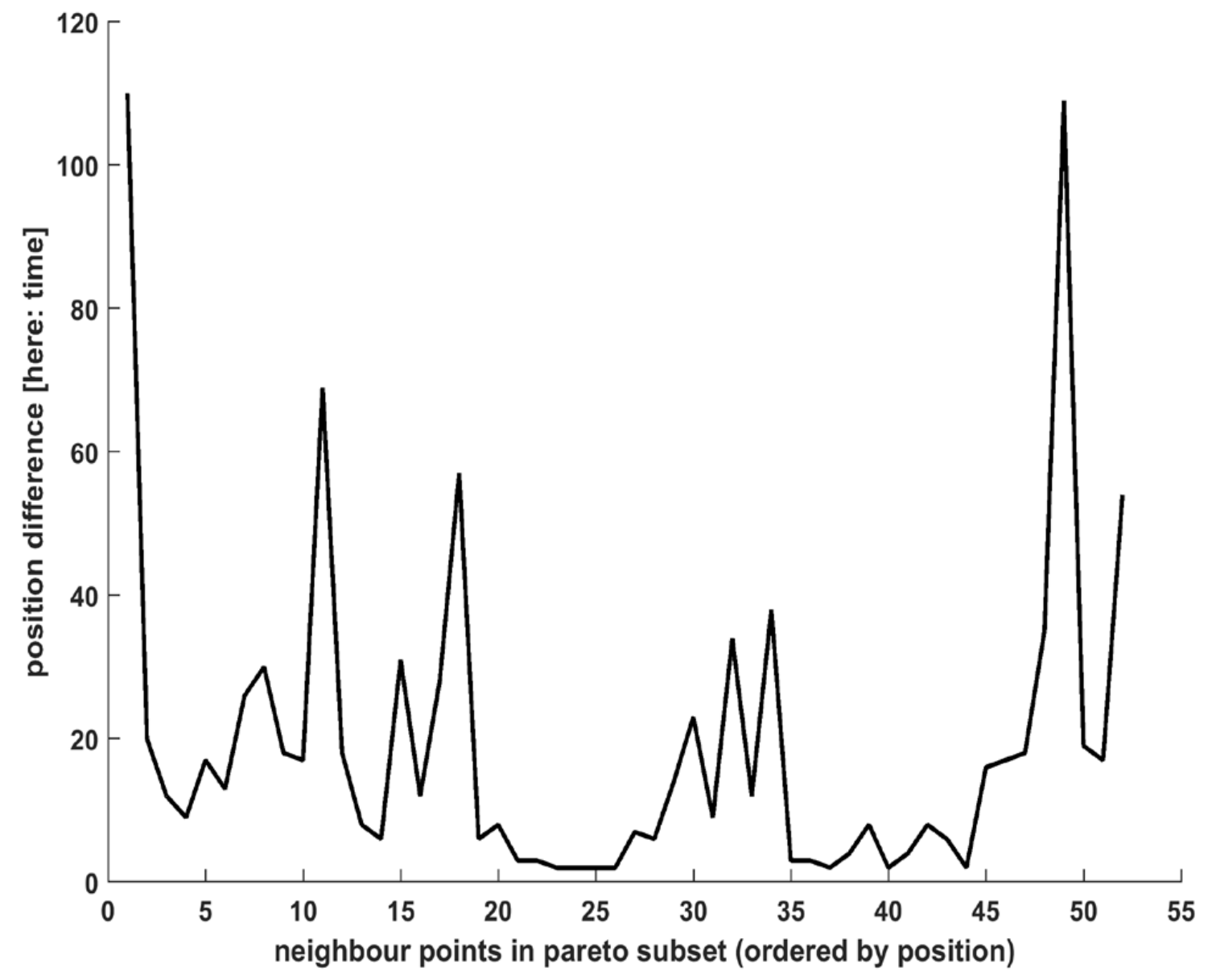

Figure 3: Position (time) difference between neighboring points of the pareto optimal subset.

3 Application

Structograms are determined for two groups of data sets: 'Sawtooth' series with increasing number of nodes (Fig. 4, A, B, and C) and different complexity (Fig. 4 B, and B1), and 'real world' data with different degrees of apparent structuredness (Fig. 5).

The Structograms of the latter (see Fig. 6 and Table 1) clearly show increasing structuredness of white noise, pink noise, temperature to discharge. For discharge, only 0.5% of the data are required to outperform the mean (not shown), and only 5,3% of the data points yield the pareto optimum!

.

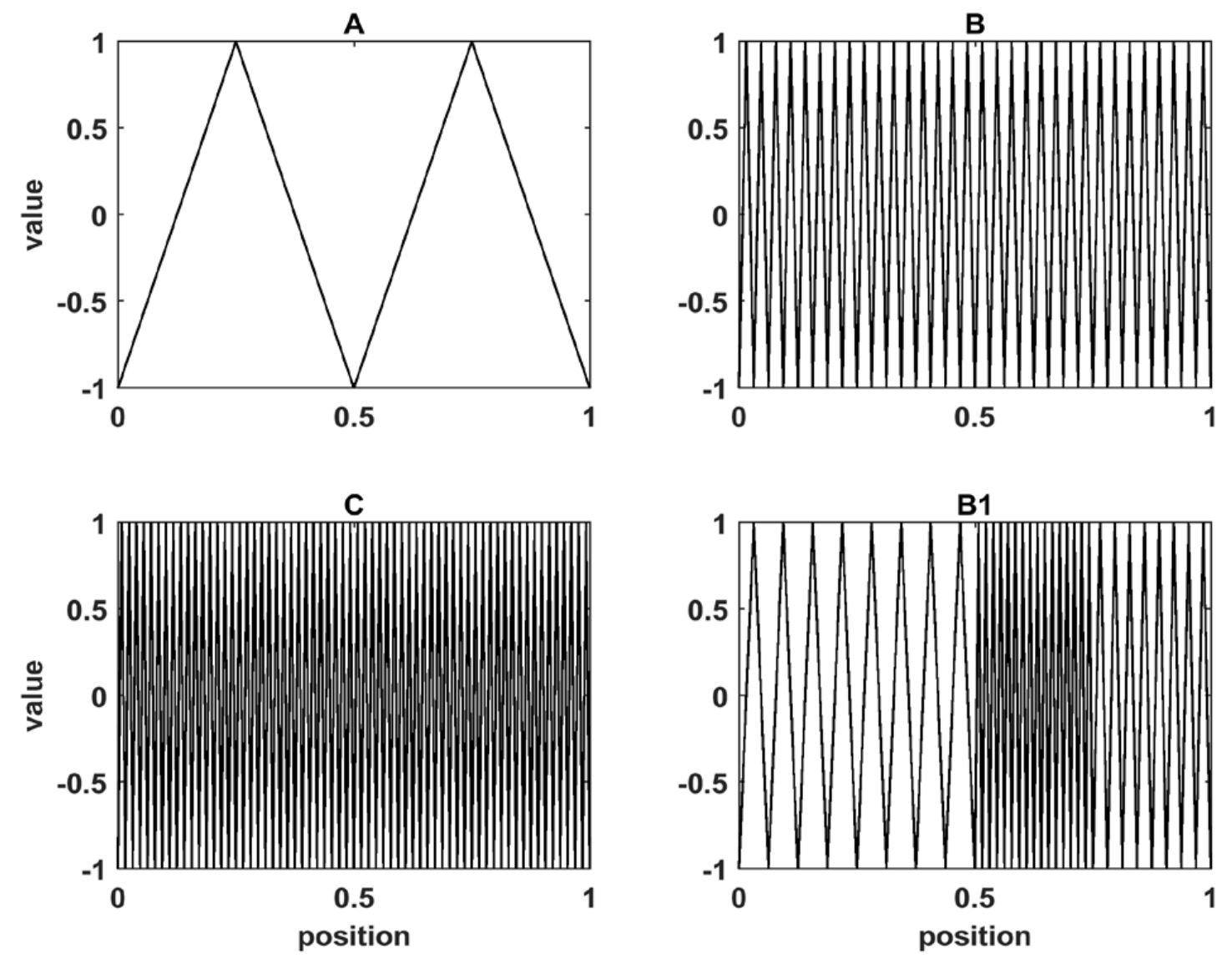

Figure 4: Sawtooth-like test data sets

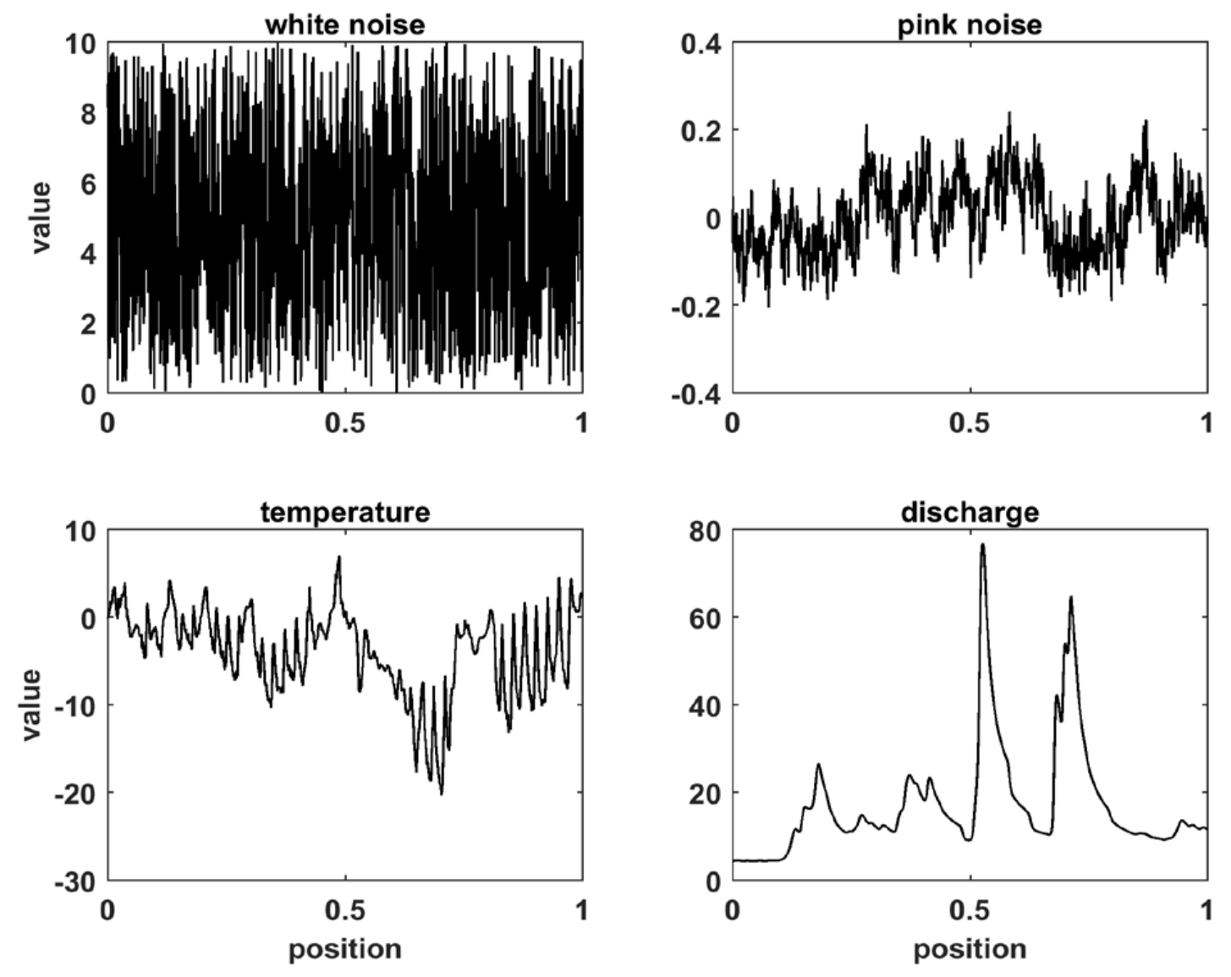

Figure 5: Real-world test data sets

Application (cont'd)

Tabel1 reveals that all sawtooth sets are highly structured, decreasing with the number of nodes. Note that both B and B1 are equally structured, but B1 has higher complexity (Table 2) due to its irregular pattern. This is in accordance with expectations.

For the real world data, the series of time differences between pareto points are shown in Fig. 7 and Table 2. White noise has the most complex pattern , discharge the least.

Note that the Structogram is invariant to additive and multiplicative transformation and the length of the data set.

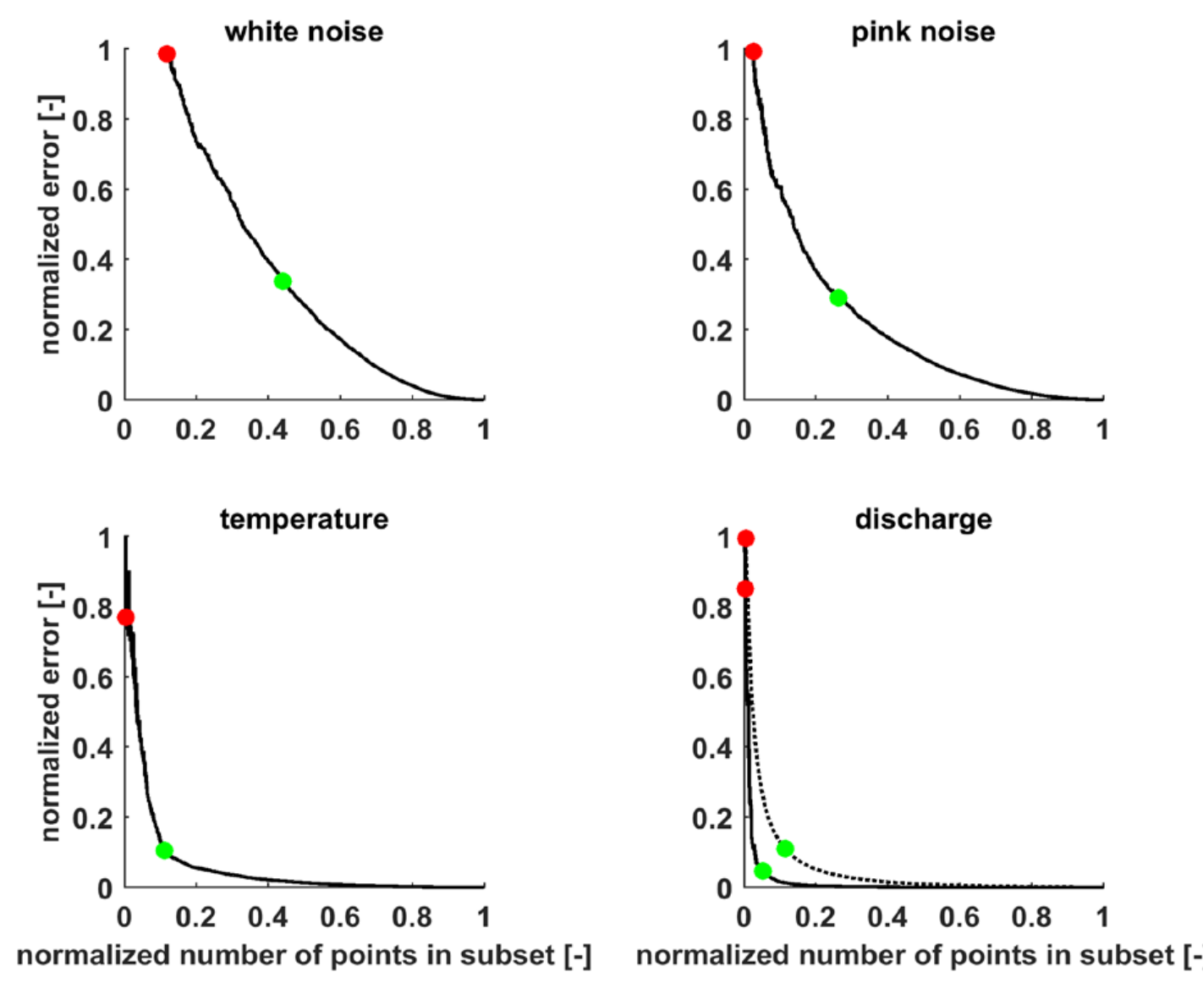

Figure 6: Structograms of real data

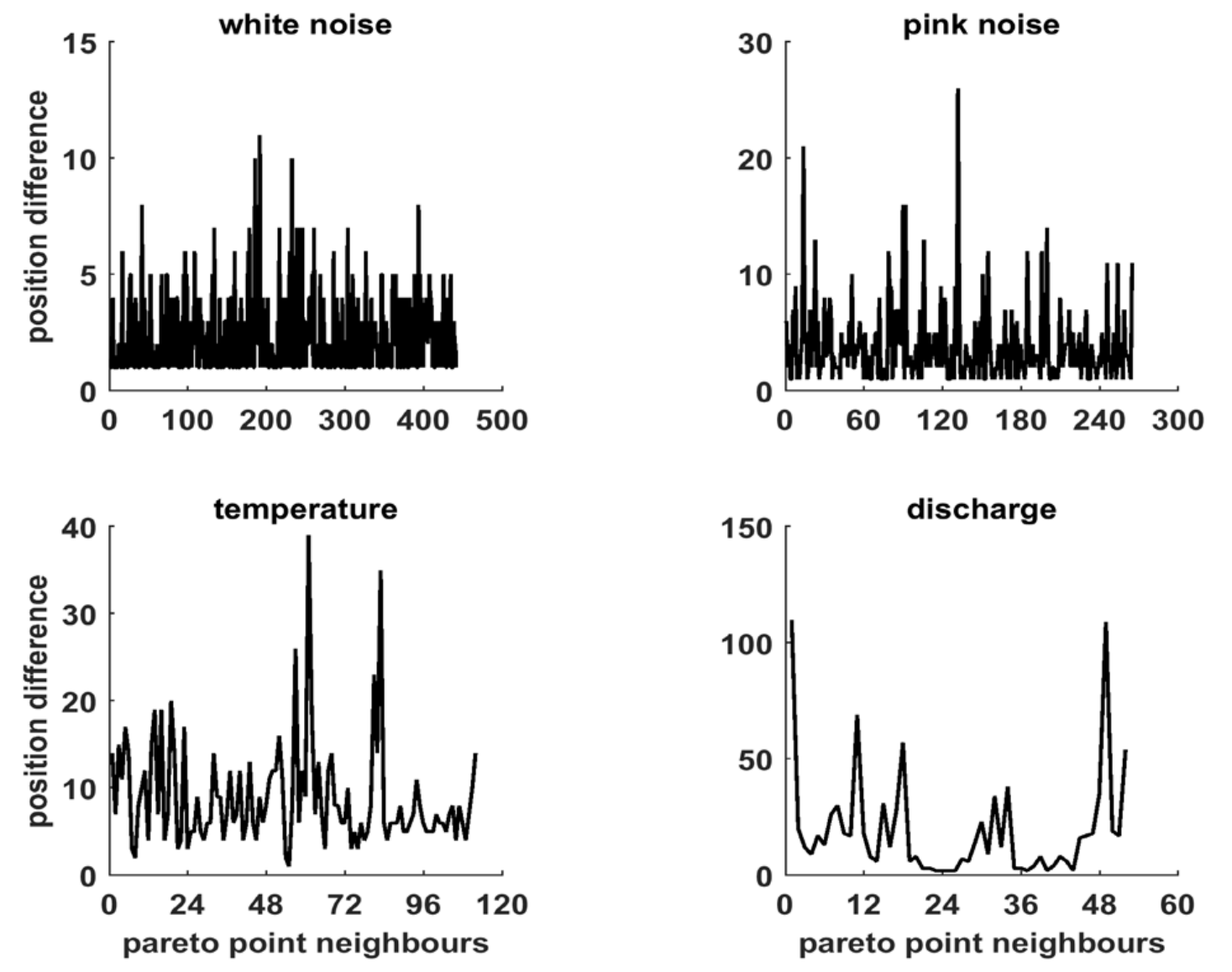

Figure 7: Time difference of pareto points

Table 1: Structogram statistics

Data set	Pareto points [%]	Structuredness [-]
White noise	44,2	0,56
Pink noise	26,4	0,74
Temperature	11,3	0,89
Discharge	5,3	0,95
Sawtooth A	0,5	0,99
Sawtooth B	6,3	0,94
Sawtooth B1	6,3	0,94
Sawtooth C	12,5	0,88

Table 2: Complexity statistics

Data set	# points [-]	Pareto points [%]	Complexity [-]
White noise	442	20,1	0,20
Pink noise	264	11,6	0,12
Temperature	113	4,1	0,04
Discharge	53	1,7	0,02
Sawtooth A	5	0,1	0,001
Sawtooth B	65	0,1	0,001
Sawtooth B1	65	0,5	0,005
Sawtooth C	128	0,1	0,001

4. Conclusions and Outlook

Applications of the Structogram include the characterization of data sets, design and evaluation of monitoring networks and interpolation. Unlike Variogram-based Kriging, the Structogram does not assume second-order stationarity. It can be applied to time series and spatial data sets of arbitrary dimension.