UFZ CENTRE FOR ENVIRONMENTAL RESEARCH LEIPZIG-HALLE

H53F-0536: A Generalization of the Local Estimator Technique

Luis E. Samaniego[1], András Bárdossy[2]

[1]UFZ-Centre for Environmental Research, Germany (luis.samaniego@ufz.de)

[2]University of Stuttgart, Germany (bardossy@iws.uni-stuttgart.de)

1. Introduction

Nearest neighbor techniques are commonly used in cluster analysis and statistics either to classify objects into a predefined number of categories or to assess the value of a predictand based on a given set of characteristics or predictors. These techniques are specially useful if the relationship between the variables is highly nonlinear. In most studies, however, the distance measure is adopted *a priori* and applied to the whole set of observations. In this study, on the contrary, a general procedure to find a metric that combines a local variance reducing technique and a linear embedding of the observation space into an appropriate Euclidean space is proposed[2].

2. Basic Definitions and Notation

System $\longrightarrow y = f(\mathbf{x}) + \varepsilon$

Data set $\longrightarrow \mathcal{D} = \{(\mathbf{x}_i, y_i) \quad i = 1, \ldots n\}$

Transformation $\longrightarrow \mathbf{u} = B[\mathbf{x}]$

Lipschitz cond. $\longrightarrow |y_i - y_j| < L d_B(i,j) \quad \forall\, i,j$

Question How to find the transformation B, so that it preserves the local continuity and is invariant with respect to changes of scale of the inputs?

Notation

y	The output of a system (a scalar or a vector).
$f(\cdot)$	A nonlinear implicit function.
$\mathbf{x}$	m-dimensional vector of inputs.
ε	Error term with mean zero and undefined distribution.
n	The sample size of the data set $\mathcal{D}$.
B	Transformation (possibly nonlinear).
$\mathbf{u}$	k-dimensional vector space ($k \leq m$).
$d_B(i,j)$	The Euclidean distance between $\mathbf{u_i} = B[\mathbf{x_i}]$ and $\mathbf{u_j} = B[\mathbf{x_j}]$.
L	A constant.
p, p^*	Threshold proportions.
$D_B(p)$	A limiting distance.
$\mathcal{N} = \lvert . \rvert$	Cardinality of the set $\lvert\{(i,j)\ ;\ d_B(i,j) < D_B(p)\}\rvert$.
N	Number of close neighbors.
λ_i	Kriging weights.
x_1	Trimmed mean slope.
x_2	Fraction of impervious cover.
x_3	Mean annual precipitation.
x_4	Mean maximum temperature in January.
x_5	Spatial variance of the precipitation.
x_6	Depth of the precipitation forecast.
$b_1, \ldots, b_7$	LANDSAT bands.

3. Method

The simplest type of transformation is linear, e.g. using a matrix:

$$\mathbf{u} = \mathbf{B}\mathbf{x}$$

B can be estimated by

$$\int_0^{p^*} G_B(p)dp \rightarrow \min$$

where

$$G_B(p) = \frac{1}{\mathcal{N}(D_B(p))} \sum_{d_B(i,j)<D_B(p)} (y_i - y_j)^2$$

$G_B(p)$ is a "local variance" function that expresses the increase of variability of the output with respect to the increase of the distance of the nearest neighbors in a nonparametric form. A solution of the objective function $G_B(p)$ (i.e. the elements of the matrix $\mathbf{B}$) can be found by Simulated Annealing[1]. The "best" dimension k of the space into which the variables x are embedded can be selected with the help of the Mallows' C_P statistic.

4. Local Estimators

Nearest neighbor

$$y = y_{i_0}$$
$$d_B(\mathbf{u}, \mathbf{u_{i_0}}) \leq d_B(\mathbf{u}, \mathbf{u_i}) \quad i = 1, \ldots, n$$

Mean of close neighbors

$$y = \frac{1}{N} \sum_{d_B(\mathbf{u},\mathbf{u_i})<D(N)} y_i$$

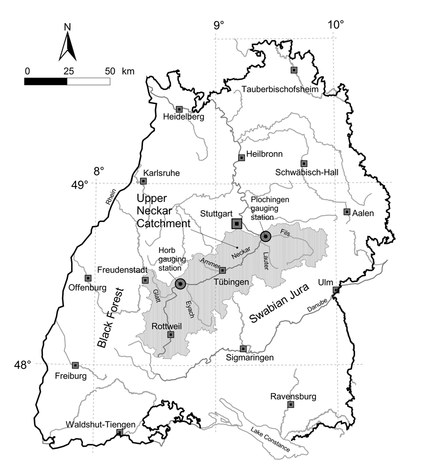
Local linear regression

$$y = a_0 + \sum_{i=1}^{k} a_i u^{(i)}$$
$$a_i \rightarrow \{(\mathbf{u_i}, y_i)\ ; d_B(\mathbf{u}, \mathbf{u_i}) < D(p_s)\}$$

Local Kriging

$$y = \frac{1}{N(p_s)} \sum_{d_B(\mathbf{u},\mathbf{u_i})<D(p_s)} \lambda_i y_i$$

5. Study Area



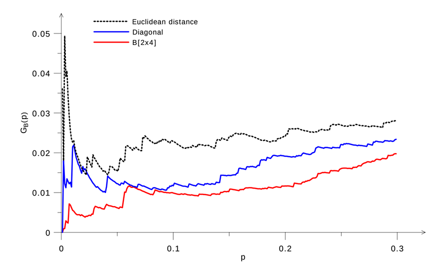
Location of the Upper Neckar Catchment

- Area: 4000 km^2.
- Elevation: ranges from 240 m to 1014 m a.s.l. with a mean of 546 m.
- Slopes: mild; 90% of its area has slopes varying from $0°$ to $15°$. In some places in the Black Forest up to $50°$.
- Climate: C_f (Köppen's notation), moist mid-latitude climates with mild winters with a mean annual precipitation of 900 mm.

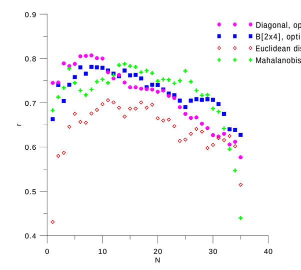
6. Results

a) Prediction of mean annual discharge (y)[2]

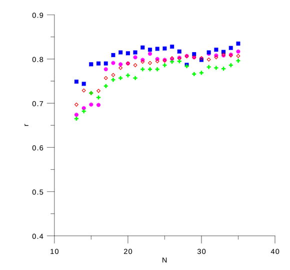
$$y_i = f(x_1, x_2, x_3, x_4) + \epsilon_i \quad i = 1, \ldots, 46$$



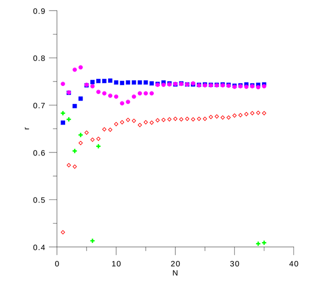
Variance functions



Mean of close neighbors



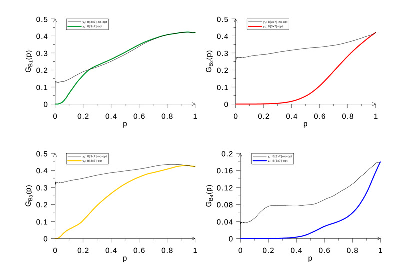
Local linear regression



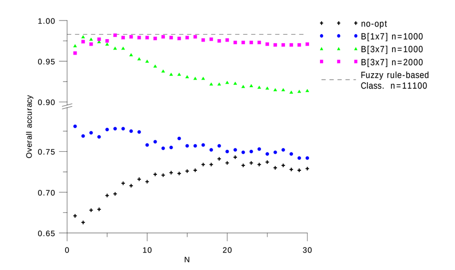
Local Kriging

b) Land cover classification (y_l)[3]

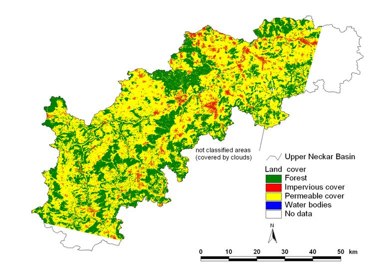
$$\{0, \ldots, y_l, \ldots, 0\}_i = f(b_1, \ldots, b_7) + \epsilon_i \quad i = 1, \ldots, 1000$$



Variance functions per Land cover class



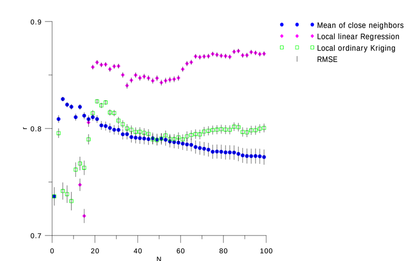
Overall accuracy w.r.t. n and $\mathbf{B}$



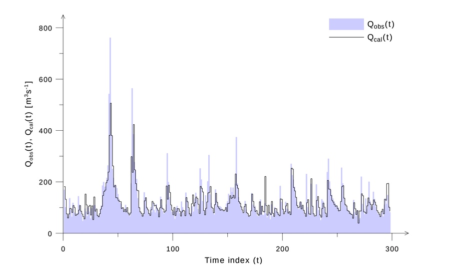
Land cover classification (1984)

c) One day flood ($\Delta Q(t)$) forecasting [2]

$$\Delta Q(t) = f\left(Q(t), \Delta Q(t-2), x_5, x_6\right) + \epsilon(t) \quad t = 1, \ldots, 586$$



Performance of the estimators



Observed vs. simulated discharges

7. Conclusions

- The optimal embedding ensures the highest degree of continuity (i.e. the "local variance" function) and it is scale invariant.
- Results show that the proposed method leads to better results than classical function fitting or the usual nearest neighbor method.
- Nonlinear embeddings might further improve this method. Further research is still needed to confirm this hypothesis.

References

[1] E. H. L. Aarts and J. Korst, *Simulated Annealing and Boltzmann Machines: A Stochastic Approach to Combinatorial Optimization and Neural Computing*. Chichester: John Wiley and Sons, 1989.

[2] A. Bárdossy, G. S. Pegram, and L. Samaniego, "Modeling data relationships with a local variance reducing technique: Applications in hydrology," *Water Resour. Res.*, vol. 41, 2005.

[3] A. Bárdossy and L. Samaniego, "Fuzzy rule-based classification of remotely sensed imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 40, no. 2, pp. 362–374, 2002.