H31E-1466: Finding an Appropriate Similarity Measure for Catchment Characterization

Luis E. Samaniego[1], András Bárdossy[2], and Karsten Schulz[1]

[1]Helmholtz - Centre for Environmental Research - UFZ, Germany (luis.samaniego@ufz.de); [2]University of Stuttgart, Germany

1. Introduction

A vast number of commonly used Euclidian and non-Euclidian similarity measures has been applied in a range of hydrological studies during the last decades, especially for catchment characterization and flood frequency analysis. All these studies have one common feature: the selection of the metric is *a priori*. In this paper, on the contrary, we propose a general procedure to find an adaptive metric that combines a local variance reducing technique and a linear (or non-linear) embedding of the observation space into an appropriate space.

2. Basic Definitions and Notation

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

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

Metric $\longrightarrow d_B(i,j)^2 = (\mathbf{u}_i - \mathbf{u}_j)\mathbf{g}(\mathbf{u}_i - \mathbf{u}_j)^{\mathrm{T}}$

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

Notation

y	The output of a system (a scalar or a vector).
$\mathbf{x}$	m-dimensional vector of inputs.
n	The sample size of the data set $\mathcal{D}$.
B	Embedding transformation (possibly nonlinear).
$\mathbf{g}$	k-dimensional metric tensor.
$\mathbf{u}$	k-dimensional vector in the embedding space ($k \leq m$).
$d_B(i,j)$	Distance between $\mathbf{u}_i = B[\mathbf{x}_i]$ and $\mathbf{u}_j = B[\mathbf{x}_j]$.
L	A constant.
p	Threshold proportion.
$D_B(p)$	A limiting distance.
$\mathcal{N}$	Cardinality of the set $\|\{(i,j)\ ;\ d_B(i,j) < D_B(p)\}\|$.
N	Number of close neighbors.
Σ	Covariance matrix, $\mathrm{cov}(\mathbf{x}_i, \mathbf{x}_j)$.
$\mathbf{I}_k$	Identity matrix.
α	Calibration coefficients.
x_1	Area.
x_2	Trimmed mean slope.
x_3	Fraction of north facing slopes.
x_4	Elevation difference (Hmax - Hmin).
x_5	Fraction of karstic formation.
x_6	Fraction of impervious cover.
x_7	Mean annual precipitation.
x_8	Mean temperature in January.
$M_l(q_i, q_j)$	Similarity measure based on a density copula $c(\cdot)$.
q_i^t	Discharge time series for basin i.
y	Runoff characteristic (e.g. $\hat{q}$, q_5).

3. Metrics

Euclidean Metric $\quad \mathbf{g} = \mathbf{I}_k = diag(1, 1, \ldots, 1)$

Mahalanobis' Metric $\quad \mathbf{g} = \Sigma^{-1}$

Riemannian Metric $\quad \mathbf{g} = [g_{ij}]$

Example $\quad [g_{ij}]$ is positive definite

$$g_{ij} = 1 + \alpha_{i,j} u_i u_j \quad \forall i = j$$

$$g_{ij} = \alpha_{i,j} u_i u_j \quad \forall i \neq j$$

4. Method

Objective Find $B[\cdot]$ and $\mathbf{g}$ so that:

- Local continuity is preserved.
- Transformation should be invariant w.r.t. the scale of the inputs.
- Shortest distance between two points (geodesics) is not necessarily a straight line.

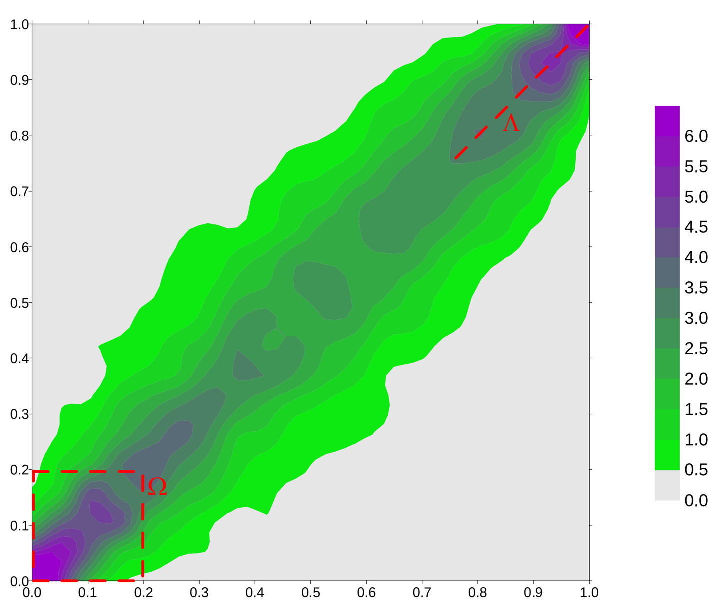
The simplest type of transformation is linear, e.g. using a matrix: $\mathbf{u} = \mathbf{Bx}$. $\mathbf{B}$ can be estimated by

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

$$G_B(p) = \frac{1}{\mathcal{N}(D_B(p))} \sum_{d_B(i,j) < D_B(p)} \Big| y(\mathbf{x}_i) - y(\mathbf{x}_j) \Big|^2$$

$$|y(\mathbf{x}_i) - y(\mathbf{x}_j)| \propto M_k(q_i, q_j)$$

$G_B(p)$ 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 [1].



Similarity Measures based on a density copula $c()$

$$C(w,v) = P\Big[F_q(q_i) < w; F_q(q_j) < v\Big] = C\Big(F_q(q_i), F_q(q_j)\Big)$$

Similarity Measures

$$M_0 = \int\int_\Omega c(q_i, q_j) d\Omega$$

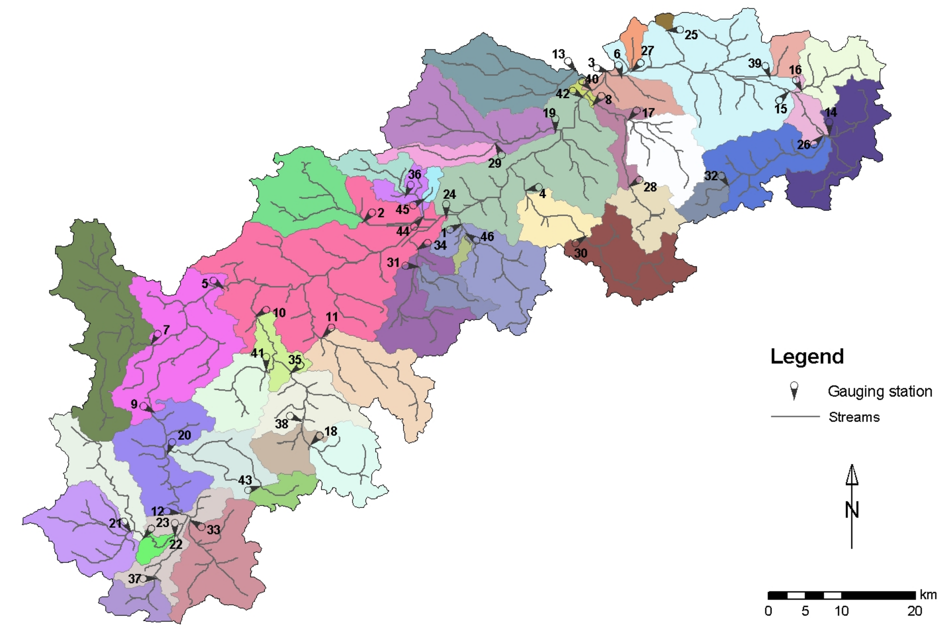
$$M_1 = \int_\Lambda \lambda c(q_i, q_j) d\lambda$$

$$c(w,v) = \frac{\partial^2 c(w,v)}{\partial w \partial v}$$

Validation: Mean of close neighbors using a runoff characteristic

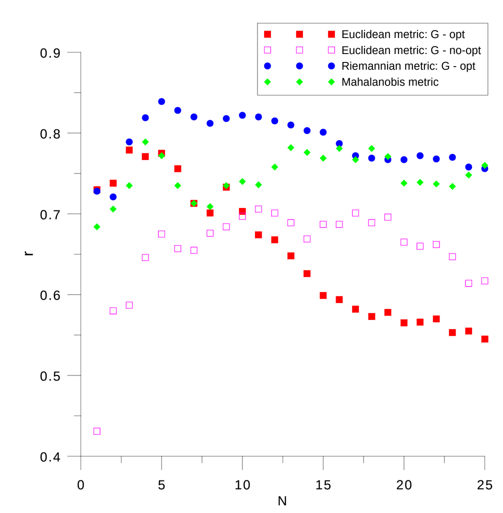
$$y = \frac{1}{N} \sum_{d_B(\mathbf{u}, \mathbf{u_i}) < D(N)} y_i$$

5. Results

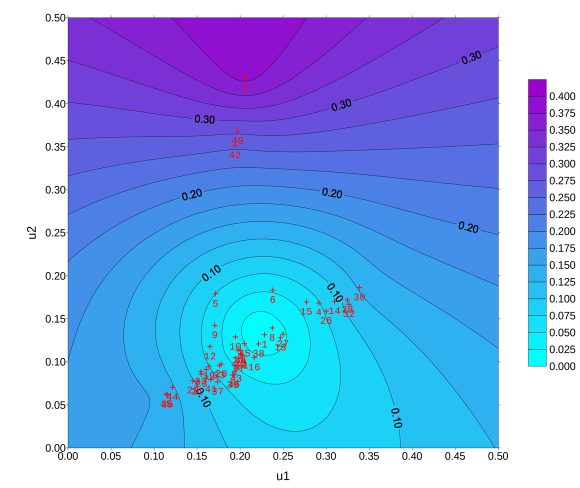


Gauging stations

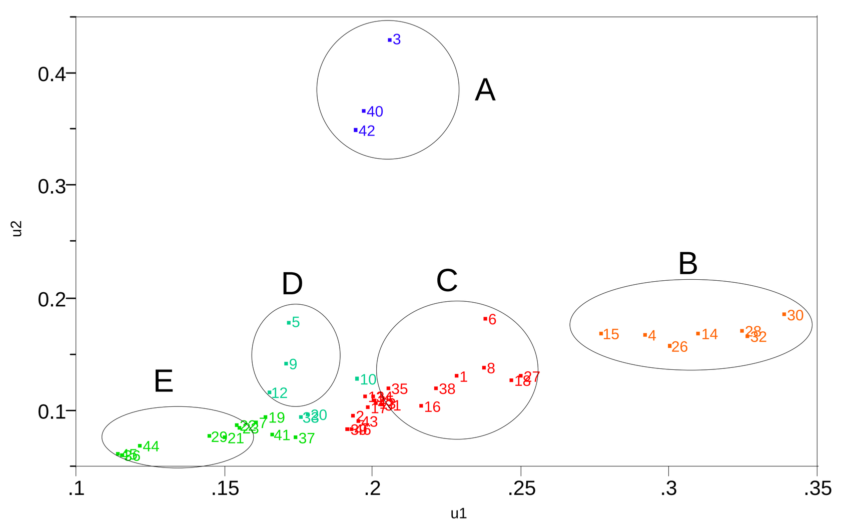
- Location: Upper Neckar Catchment, Germany
- Area: 4000 km^2.
- Elevation: ranges from 240 m to 1014 m a.s.l. with a mean of 546 m.
- Slopes: mild; 90% $0°$ to $15°$.
- Precip.: $\approx$ 900 mm/yr.

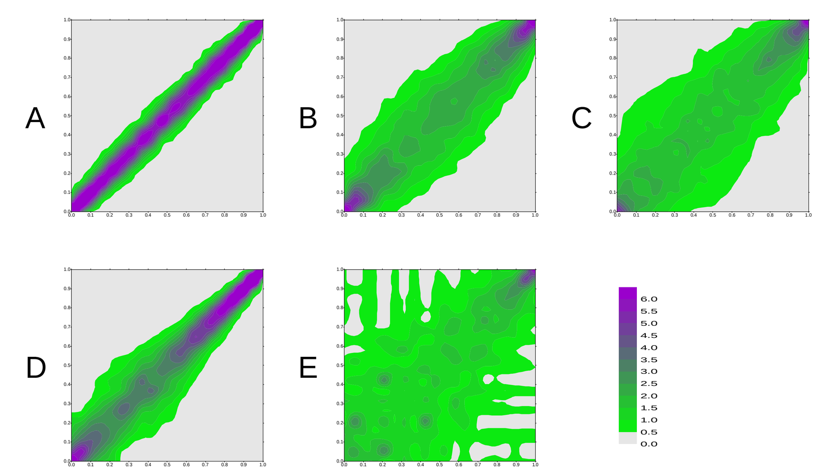


Metric performance



Contours of constant (shortest) distance from basin 1 to all other basins





Left: Clusters obtained with M_0. Right: Typical pairwise runoff density copulas within a given cluster

6. Conclusions

- Use of an embedding space and adaptive metric performed much better than *a priori* selected standard metrics.
- Similarity measures based on density copulas lead to robust classifications. Validation with several runoff characteristics (y) gives $r > 0.7$.

References

[1] 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.

HELMHOLTZ CENTRE FOR ENVIRONMENTAL RESEARCH – UFZ