IDENTIFYING HOMOGENEOUS REGIONS USING STATISTICAL DEPTH FUNCTION

Hussein Wazneh*[1], Fateh Chebana[1] & Taha Ouarda[2]

[1] *INRS-ETE, University of Quebec,* [2] *Masdar Institute of science and technology*

5rd STAHY International Workshop, Abu Dhabi

I. Introduction

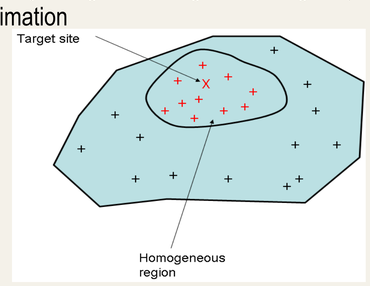
Extreme events (floods, droughts, storms) have serious social and economic consequences.

Local frequency analysis is a favourite tool to predict the frequency of an extreme event when sufficient hydrological information is available.

Regional frequency analysis (RFA) is used to estimate extreme events at ungauged sites or sites with short records.

RFA is composed of 2 main steps:

- Delineation of hydrological homogeneous region (**DRH**)
- Regional estimation



Traditional approaches to DRH **:**

- Canonical correlation analysis, Region of influence
- **Hierarchical Clustering (HC)**

Motivation

Drawbacks related to the traditional HC approach:

- **Based on distance measures** (e.g., Ward or linkage)
- **Uses non robust statistics** (e.g., k-means) :
- **Sensitive to noise and to outliers**
- Require a **pre-selection of the number of sub-regions**

Aims of this study:

- Propose robust approach for DRH based on the notion of depth function
- Make the delineation step automatic and objective

II. Tools

Spatial Depth function

A depth function is a statistical notion introduced to extend order in multidimensionnal spaces.
Especially, to define the median of multivariate sample.

Spatial depth : The spatial depth of point x is the amount of probability mass needed at x to make it the multivariate median (spatial median) of the data :

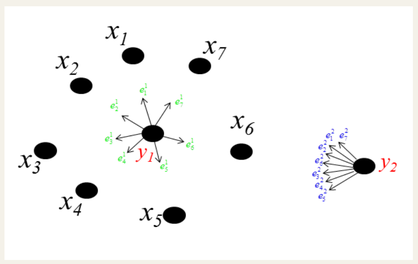
$$SPD\left(x,\hat{F}_X\right)=1-\frac{1}{n}\left\|\sum_{i=1}^{n}\frac{x-x_i}{\|x-x_i\|}\right\|,\quad x\in\Re^d$$

Example :

$X=\{x_1,....,x_7\}$

$SPD(y_1,X)\simeq 1$

$SPD(y_2,X)\simeq 0$



Index-flood model

The index-flood model estimates flood quantile corresponding to a nonexceedance probability t through the expression :

$$Q_{i_0}(t)=\mu_{i_0}q_t\left(\theta^R\right)\quad ;0<t<1 \text{ and } \hat{\theta}_l^R=\sum_{h=1}^{N'}\frac{n_h}{\sum_{h=1}^{N'}n_h}\hat{\theta}_l^h,\quad l=1,..L$$

i_0 : Identifier of the target site; $q_t(.)$: Growth curve function

μ_{i_0} : Index flood; $\hat{\theta}_l^h$: l^{th} parameter obtained from the h^{th} gauged site

θ^R: Vector with L components of the regional growth curve parameters

N': Number of sites in the homogenous sub-region that contain the site i_0

Depth index-flood model

The depth index-flood model estimates flood quantile using the same expression of index-flood model. However, the vector of regional parameter θ^R is estimated by :

$$\left(\hat{\theta}^R\right)_D=\sum_{h=1}^{N'}\omega_h\theta_l^h,\quad l=1,...,L \text{ and } \omega_h=\varphi\left[Depth\left(\hat{\theta}^h;\Theta\right)\right];\ h=1,...,N'$$

φ : Increasing weight function (e.g., Gompertz, logistic,..);

Depth : Depth function (e.g., Mahalanobis, Tukey,..)

$\Theta=\{\hat{\theta}^1,....,\hat{\theta}^{N'}\}$: Set contain the vectors of at-site parameter

III. DRH using spatial depth function

Particular definition

$X_r=(X_{r1},...,X_{rp})$: p selected attributes for site r;

K : number of homogeneous sub-regions;

N: number of sites in the data set;

$I(k)=\{X_r; r\in \text{sub-region } k; r=1,..,N\};\ k=1,..K$

We define the:

- Within sub-region depth of site r: $D_r^w=SPD\left(X_r,I(k)\right)$

D_r^w quantifies how central the site r is with respect to its sub-region k.

- Between sub-region depth of site r: $D_r^b=\max_{\substack{l=1,..K\\ l\neq k}}\left[SPD\left(X_r,I(l)\right)\right]$

D_r^b is the maximum depth of site r with respect to the sub-region to which it does not belong.

-Deviance sub-region depth of site r: $\text{DeD}_r=D_r^w-D_r^b$

D-clustering criterion

- $\text{DeD}_r>0 \Rightarrow$ site r is well classified in its sub-region
- $\text{DeD}_r\simeq 0 \Rightarrow$ site r lies between two sub-region
- $\text{DeD}_r<0 \Rightarrow$ site r is placed in the wrong sub-region

Criterion: in the D-clustering approach, each site r is assigned to the sub-region that maximizes its DeD_r

Example :

$X_r=(AREA_r,(H_m)_r)$ with $H_m=$ mean elevation; $K=3$; $X_1,X_2\in I(1)$

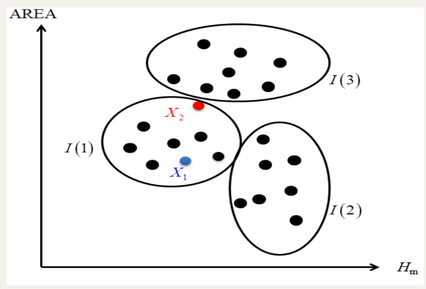
$D_1^w=SPD\left(X_1,I(1)\right)$ and $D_2^w=SPD\left(X_2,I(1)\right)$;

Suppose that: $SPD\left(X_1,I(2)\right)>SPD\left(X_1,I(1)\right)>SPD\left(X_1,I(3)\right)$ and

$SPD\left(X_2,I(3)\right)>SPD\left(X_2,I(1)\right)>SPD\left(X_2,I(2)\right)$;

In this case $D_1^b=SPD\left(X_1,I(2)\right)$ and $D_2^b=SPD\left(X_2,I(3)\right)$

Using the D-clustering criterion, these two sites must modify their sub-regions. In fact, site 1 must be assigned to sub-region 2 and site 2 must be assigned to sub-region 3.



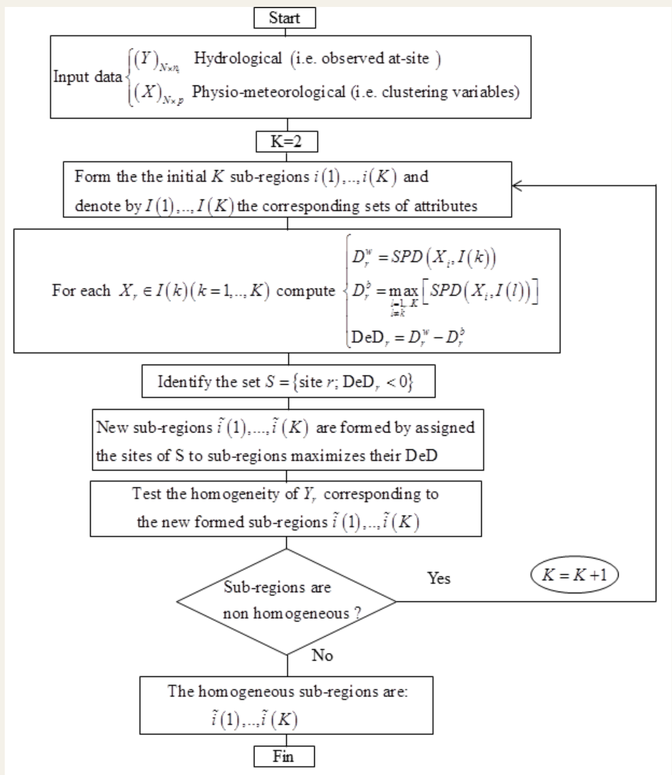
D-clustering algorithm

To find the optimal homogeneous sub-regions, the procedure is composed of three main steps :

- Initialization: Use a traditional clustering approach to form the initial K sub-regions (we start with K=2, see Algorithm below);

- Modifying: Modify the initial position of sites r using the D-clustering criterion (i.e. each site r is reassigned to the sub-region that maximizes its DeD);

- Homogeneity: Test the homogeneity of the final sub-regions obtained after the modification step. If the obtained sub-regions are not homogeneous go to the initial step with K=K+1.



IV. Case study and results

Region: Piemonte, Valle d'Aosta and Liguria, three geographical regions in the North-West of Italy.

Data: Annual maximum peak flood data of 47 stream flow gauging sites.

Clustering variables: coordinates in the UTM system of the catchment centroids (X_{bar} and Y_{bar}) and the mean elevation (H_m) of the drainage basin.

Aims:

1) Delineate the homogeneous sub-regions using Ward and D-clustering approaches;

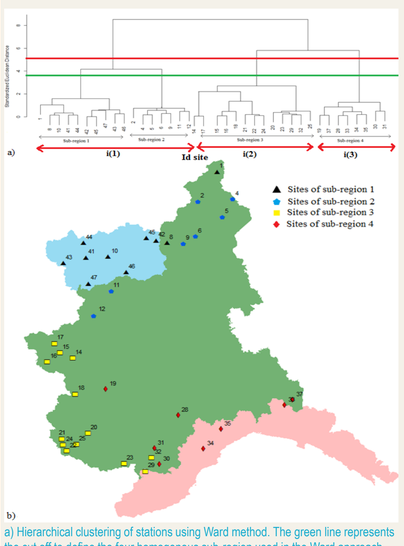
2) Estimates the flood quantile for different non-exceedence probability and using the obtained sub-region by the different DRH approaches.

Performance criteria:

Since the delineation step affects the results of estimation of flood quantiles, two categories of performance criteria are used in this study:

Criteria assess the clustering approaches	criteria assess the estimation results
H: Heterogeneity measure	RB: Relative bias
SIL: Silhouette of sub-region	RRMSE: Relative root mean square error
GSIL: Global silhouette	

Selected results



a) Hierarchical clustering of stations using Ward method. The green line represents the cut off to define the four homogenous sub-region used in the Ward approach. The reed line represent the cut off to define the three initial sub-region used for the D-clustering approach. b) Geographical location of sites of four homogeneous sub-regions formed using Ward method.

Id site	Initial sub-region	DeD	Final sub-region
6	1	-0.10	3
11	1	-0.12	2
12	1	-0.21	2
14	2	-0.15	3
17	2	-0.13	3
20	2	-0.17	3
23	2	-0.21	3
29	2	-0.15	3
32	2	-0.13	3

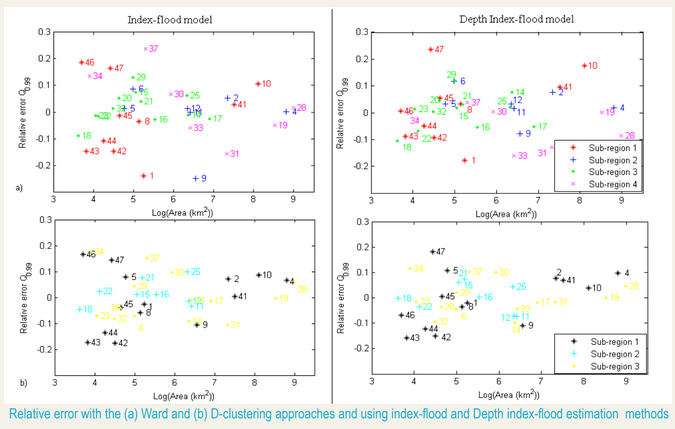
. Sites that have negative deviance depth "DeD". These sites are moved from their initial sub-regions in the D-clustering approach.

Delineation approach	Sub-region	Number of sites	H	SIL	GSIL
Ward	1	10	1.93	0.41	0.43
	2	7	1.96	0.39	
	3	13	-0.3	0.41	
	4	8	0.5	0.51	
D-clustering	1	14	1.03	0.44	0.47
	2	9	-0.6	0.49	
	3	15	0.014	0.48	

Heterogeneity measures H, the silhouette SIL and the global silhouette $GSIL$ using traditional Ward and D-clustering approaches.

Delineation approach	Estimation method	0.9		0.99		0.995		0.999	
		RB	RRMSE	RB	RRMSE	RB	RRMSE	RB	RRMSE
Ward	Index-flood	0.21	4.43	0.23	10.70	0.18	12.70	-0.05	17.33
	Depth-based index-flood	0.18	4.19	0.32	8.75	0.35	10.65	0.86	14.43
D-clustering	Index-flood	-0.78	3.25	0.18	9.05	0.32	9.40	-0.12	11.54
	Depth-based index-flood	-0.42	3.59	-0.21	8.00	0.44	8.77	0.07	10.17

Quantile estimation results in % using the index-flood and Depth-based index-flood method for the studied region using D-clustering and Ward approaches



Relative error with the (a) Ward and (b) D-clustering approaches and using index-flood and Depth index-flood estimation methods

V. Conclusion

- **The proposed D-clustering approach is robust.**
- **This algorithm based-procedure automates the optimal choice of the sub-region with respect to the homogeneity criterion.**
- **It allows the DRH to be more practical and usable without the user's subjective intervention.**
- **The study of the three regions shows that the proposed D-clustering approach leads:**
  to more homogeneous sub-regions in terms of *H* heterogeneity measure
  to more efficient quantile estimations in terms of relative bias and relative root mean square error
  than those obtained by the traditional DRH approach.

Contact Information

Hussein Wazneh, PhD student

490 rue de la couronne
G1K 9A9, Québec, Canada

Tel: 418 654 2430#4468
Email: hussein.wazneh@ete.inrs.ca