Optimal temporal distribution curves for the classification of heavy precipitation using hierarchical clustering on principal components

Konstantinos Vantas[1,*], Epaminondas Sidiropoulos[1], Marios Vafeiadis [1]

[1]Faculty of Engineering Aristotle University of Thessaloniki GR- 54124 Thessaloniki, Macedonia, Greece

[*]Corresponding author:

kon.vantas@gmail.com, tel: +30 24670 24804

ABSTRACT

A novel method that utilizes a combination of statistical and clustering techniques is presented in order to classify statistically independent heavy rainstorm events and create a limited number of representative intra-storm temporal distribution curves. These curves represent the centers of many dimensionless cumulative rainstorm events and express the temporal distribution patterns in a probabilistic way. The whole process includes the necessary steps from importing raw precipitation time series data to producing the initially unknown optimal number of representative curves. These hyetographs can be used for stochastic simulation, water resources planning, water quality assessment and global change studying. The present type of analysis is fully unsupervised, as no empirical knowledge of local rainfalls is implicated or any arbitrary introduction of quartiles for grouping as is the case in the pertinent literature. It replaces the traditional Huff's method by utilizing modern machine learning techniques, thus being clearly data driven and more rational. An example using data from a Greek Water Division illustrates that the proposed method produces clusters with superior internal structure and temporal distribution curves that are not coming from the same distribution, in contrast to the results using the established Huff's curves classification.

Keywords: rainfall temporal distribution; design hyetographs; optimal number of clusters; unsupervised learning; hierarchical clustering; principal components analysis; cluster validity; Huff's curves

1. Introduction

Knowledge about the temporal distribution of rainfall is essential in current methods of water resources management such as drainage design, erosion control, water quality assessment and global change studies. A typical methodology includes the determination of total duration and height of rainfall and disaggregation of this height using a temporal pattern that represents the expected internal rainfall structure, the design hyetograph (DH). A study (Veneziano & Villani, 1999) provided categorization of methods for the production of DHs, distinguishing four types. The first two methods are based on Intensity-Duration-Frequency curves, the third method is based on standardized profiles derived from rainfall records and the last one relies on stochastic rainfall models via simulation. The first three methods are used more frequently in practice.

Huff (1967) presented a probabilistic method, in which storm data are classified using the quartile where the maximum intensity occurs. In this procedure, rainstorms are extracted and transformed to dimensionless form using the normalized cumulative precipitation as a function of the normalized rainstorm duration. More details about the development and utility of Huff's curves in disaggregation and stochastic simulation can be found in the literature (J. Bonta, 2004a, b, J. Bonta & Rao, 1987, J. V. Bonta & Shahalam, 2003). The well-known two-sample Kolmogorov-Smirnov test (KS) and chi-square test has been used to indicate whether there are statistically significant differences between two sets of Huff's curves (J. V. Bonta & Shahalam, 2003; Williams-Sether *et al.*, 2004). Huff's categorization makes the assumption that the rainstorms within a quartile are more similar to others that belong to a different one, although this grouping has been criticized as artificial without physical meaning (Koutsoyiannis, 1994). Recently, an improvement on Huff curves was proposed by separately describing the rising and falling limbs of normalized rainstorms (Pan *et al.*, 2017). Also, Bezak *et al.* (2018) recommend the use of Huff's curves for the selection of DH in hydraulic flood modelling. Despite the dispute, the National Oceanic and Atmospheric Administration provides temporal distributions similar to Huff's in the Precipitation-Frequency Atlas of the United States (Perica *et al.*, 2012) and in a number of different regions or countries

Huff's curves were used (Azli & Rao, 2010, Guo *et al.*, 2001, Loukas & Quick, 1996, Yin *et al.*, 2016, Zeimetz *et al.*, 2018). Finally, the concept of cumulative mass curves were used in conjunction with climate models to examine future changes in storm properties (Jiang *et al.*, 2016). The present paper presents a more rational and fully unsupervised classification of intra-storm temporal patterns of rainfall based on modern machine learning methods.

Learning algorithms fall into one of the categories of supervised, reinforcement and unsupervised learning (Abu-Mostafa et al., 2012). The problem of determining rainfall intra-storm temporal distribution patterns, or to group these data into meaningful clusters, when there is no output information, is one of unsupervised learning. A very large number of these algorithms exist in the literature, a classification of them can be found in Sheikholeslami et al. (1998), and the most common ones used in practice are k-means (Hartigan & Wong, 1979, MacQueen, 1967) and Hierarchical Clustering, (HC; Ward, 1963). Also, a large number of methods can be found for the evaluation of the results of clustering analysis, a task termed as the cluster validity. The validity criteria are categorized as follows (Theodoridis and Koutroumbas, 2009): a) external, where the results of clustering are compared with a priori known results, b) internal, where only the results of clustering from an algorithm are used and c) relative, where results from different clustering methods are used.

The optimal number of clusters, which in most cases is unknown, is a major issue in unsupervised learning, because different algorithms or even different parameters for the same algorithm lead to different clusters of data. A number of methods for the estimation of the optimal number of clusters, based on the relative cluster validity, can be found in (Milligan & Cooper, 1985). Feng & Hamerly (2007) utilized the univariate KS test and the Gaussian mixture model to learn the numbers of clusters in data. A comprehensive list of 30 different indices can be found in (Charrad *et al.*, 2014) and two recent papers (Zambelli, 2016, Zhou *et al.*, 2017) provide new methods and indices for determining the optimal number of clusters based on HC.

Applications of learning algorithms using hydro-meteorological data, in general, has been dealt with in the literature, in terms of supervised learning, such as estimating rainfall erosivity values (Vantas & Sidiropoulos, 2017). The use of unsupervised methods and the assessment of their validity in relation to the special issue of temporal distribution of rainfall are scarce and only recently such methods are appearing in the literature. Self-organized maps (SOM) have been applied to a small data-set to estimate design storms (Lin & Wu, 2007), wavelet transform and SOM were used to cluster spatial–temporal monthly precipitation data (Hsu & Li, 2010) and k-means clustering was used to create a predefined number of rainfall patterns (Nojumuddin & Yusop, 2015). At a recent time, Vantas, Sidiropoulos, & Loukas, (2019) utilized HC to identify regions that have similar temporal distribution of rainfall erosivity density and Vantas, Sidiropoulos, & Vafeiadis, (2019) used a data driven approach for the temporal classification of heavy rainfall using SOM.

This paper aims to present an original, controlled, fully reproducible, unsupervised method that produces automatically the optimal number of temporal distribution curves using precipitation records. This method comprises the following steps: a) Raw precipitation data are imported, b) independent rainstorms are extracted, c) Dimensionless Cumulative Hyetographs (DCH) are compiled, d) a hierarchical algorithm is applied that produces a set of clustering results, e) the optimal number of clusters is determined by a customized cluster validity method based on the two-sample Kolmogorov-Smirnov test and f) the rainstorm records are represented in a probabilistic way using a limited number of temporal distribution curves. An earlier, shorter, version, without the extensive treatment of the rainfall classification problem, was presented in the International Conference "Protection and Restoration of the Environment XIV" (Vantas *et al.*, 2018). The emphasis in this paper is on the method and the study case represents an instance of the general problem. Thus, the various analytical aspects are expounded in more detail and, furthermore, cluster tendency and the determination of the optimal number of clusters are given an additional treatment.

2. Materials and methods

The methodology that was applied in the study is presented in Figure 1 as a flowchart. High frequency precipitation data were imported and independent rainstorms were extracted, dimensionless cumulative mass curves were compiled, clustering validation, the proposed Algorithm and the Huff's classification was used. Last but not least, a comparison between the previous methods was made.

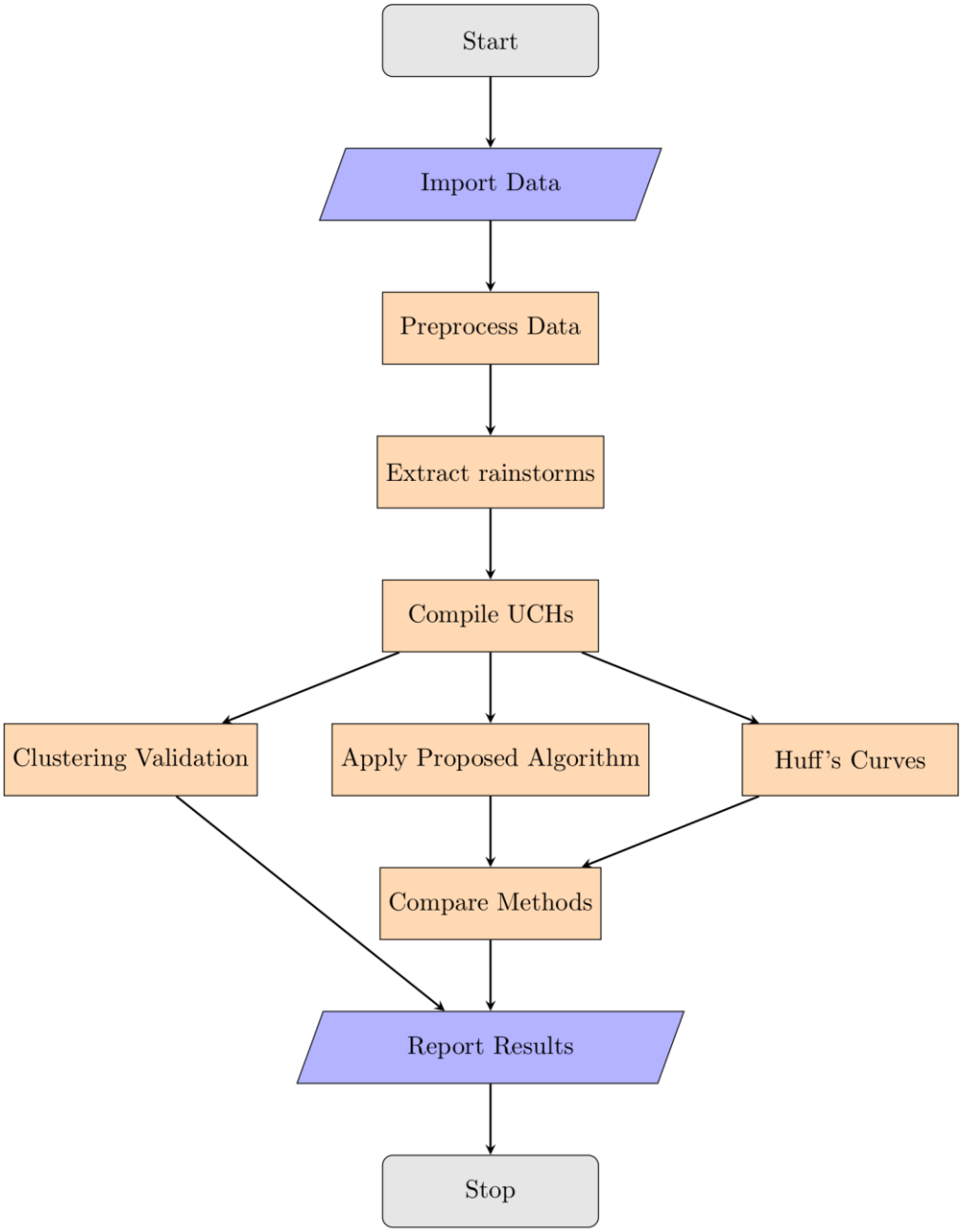


Figure 1. Flowchart of the applied methodology.

2.1. Data Acquisition and Processing

The study region, located to the north-east Greece (Figure 2), extends to an area of 11,243 km^2 that covers the Water Division of Thrace. It is delimited by the boundaries of Greece, Bulgaria and Turkey on the north and east, by the Thracian Sea on the south and by the watershed of Nestos River on the west. The climate is predominantly Mediterranean and annual rainfall ranges from 500 mm in coastal and insular areas to 1000 mm in the northern mountainous areas (Ministry of Environment and Energy, 2013).

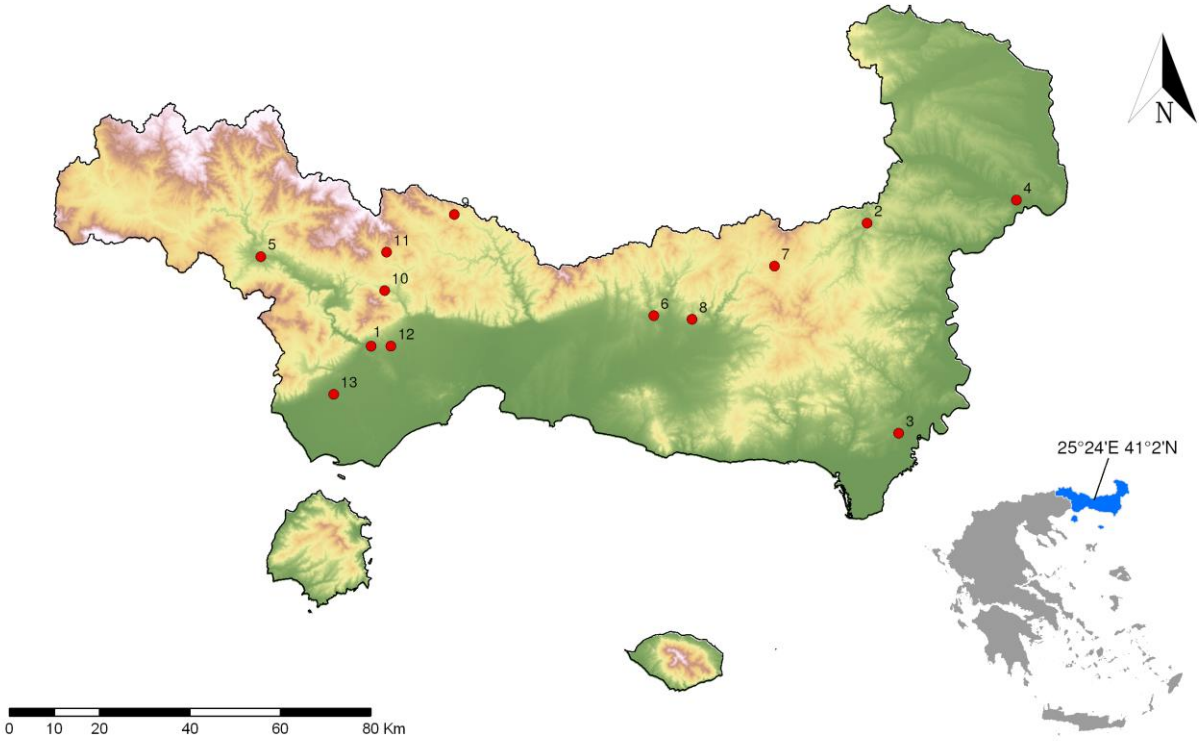


Figure 2. Location of the study area and the 13 meteorological stations

The data utilized in the analysis (Table 1) were taken from the Greek National Bank of Hydrological and Meteorological Information (Vantas, 2018a), measured at 13 meteorological stations. The data coverage (i.e. the percentage of non-missing values) was 37% on average and the time series comprised a total of 413 years of pluviograph records with a time step of 30 minutes for the time period from 1956 to 1997. The time series rainfall records were checked for consistency and cleared from errors.

Table 1. Meteorological stations information. ID is an abbreviation for the station ID as reported in the Greek National Bank of Hydrological and Meteorological Information, Lon for longitude, Lat for latitude, El for elevation, L for time-series length and MCV for mean annual data coverage per station.

	ID	Lon (°)	Lat (°)	El (m)	From	To	L (years)	MCV (%)
1	200249	24.79	41.09	75	1956	1997	41	62%
2	200259	26.10	41.32	116	1973	1997	24	63%
3	200260	26.17	40.90	43	1962	1997	35	56%
4	200263	26.50	41.35	25	1955	1996	41	62%
5	200311	24.50	41.27	122	1960	1996	36	65%
6	500250	25.53	41.14	120	1965	1996	31	21%
7	500251	25.86	41.23	700	1965	1996	31	20%
8	500253	25.64	41.13	70	1965	1996	31	25%
9	500262	25.01	41.35	440	1965	1996	31	21%
10	500265	24.83	41.20	308	1965	1996	31	26%
11	500267	24.83	41.27	656	1965	1996	31	18%
12	500272	24.84	41.09	65	1968	1992	24	21%
13	500273	24.69	40.99	15	1966	1992	26	16%

2.2. Optimal intra-storm temporal distributions curves algorithm

The unsupervised method that creates the optimal number of distribution curves utilizes a hierarchical tree and a top-down iterative procedure (Algorithm 1). A necessary step prior to the construction of Huff's curves is the extraction of individual rainstorm events from precipitation time series. Huff used a six-hour fixed Critical time Duration (CD) of no precipitation to separate these events, and many researchers followed the same approach (Azli & Rao, 2010, Dolšak *et al.*, 2016, Loukas & Quick, 1996, Williams-Sether *et al.*, 2004), although Bonta, (2001) showed that CD has seasonal variability.

Algorithm 1: Optimal number of clusters

Input: Stations' precipitation time series $\boldsymbol{P_i}$ where $i = 1, \ldots, k$; monthly minimum dry time period duration $MDPD$; significance level $\alpha = 0.05$

Use $MDPD$ to extract a set of independent rainstorms $\boldsymbol{R}$;
Compile Dimensionless Cumulative Hyetographs' matrix $\boldsymbol{U}$ using $\boldsymbol{R}$;
Apply Principal Components Analysis on $\boldsymbol{U}$;
Use Hierarchical Clustering on the first l Principal Components that explain the 99.5% of total variance of $\boldsymbol{U}$ and get the tree-based representation of the DCHs;
while *all p-values* $< \alpha$ **do**
    moving down the tree cut into q different clusters $q = 1, ..., m$;
    calculate the mean values $\bar{x}_q$ of the DCHs that belong to cluster q;
    for all $\bar{x}_q$ obtain the Kolmogorov–Smirnov two sample test, p-values;
    adjust the obtained p-values using Benjamini and Hochberg method;

Result: optimal number of clusters q_{opt} and intra-storm distribution curves $\bar{x}_{opt}$

In the proposed Algorithm a Poisson process hypothesis is assumed for the division of the precipitation time series to statistically-independent rainstorm events, in which: a) the events'

interarrival times t_a that come from the same month are distributed exponentially and b) the events are separated by a monthly, constant, minimum dry period duration of no precipitation, MDPD. The probability density function of t_a is (Restrepo-Posada & Eagleson, 1982):

$$f(t_a) = \omega \cdot e^{-\omega \cdot t_\alpha}, \quad t_a \geq 0 \tag{1}$$

where ω is the average storm arrival rate and:

$$t_a = t_r + t_b \tag{2}$$

where t_r is the storm duration and t_b is the dry time between rainstorms. More details about the specific implementation of the method can be found in Vantas et al. (2018).

The general approach for the development of Dimensionless Cumulative Hyetographs (DCH) given by Bonta (2004a) is followed and only the events in $\boldsymbol{R}$ with duration greater than 3 hours and cumulative rainfall greater than 12.7 mm are used in the analysis. The hyetographs of the rainstorms that meet these criteria are transformed to dimensionless form in which a) the cumulative rainfall expresses the percentage of total rainstorm height and b) the time expresses the percentage of the rainstorm duration:

$$p_i = \frac{h_i}{H} \tag{3}$$

where p_i is the cumulative dimensionless precipitation height, at time i, h_i is the cumulative precipitation height at time i and H is the total precipitation height;

$$d_i = \frac{t_i}{D} \tag{4}$$

where d_i is the cumulative dimensionless duration at time i, t_i is the cumulative duration at time i and D is the total rainstorm duration.

Since the DCHs' vectors in this form have variable length, linear interpolation is applied to compute the dimensionless cumulative rainfall for every 1% of dimensionless time values. Finally, a matrix of DCHs, $\boldsymbol{U}$, is produced with the values of dimensionless cumulative rainfall, in which every row represents a DCH and every column the dimensionless time values.

On the grounds that the time variables (i.e. the columns of the $\boldsymbol{U}$ matrix) are highly correlated, Principal Component Analysis (PCA, Pearson, 1901) is applied to reduce the dimensionality of the data to a few dimensions. The number of dimensions to retain is determined using the proportion of total variance of the data explained (Jolliffe, 1986). In this analysis this level is set to 99.5%, to ensure that almost all the information from DCHs will be preserved.

The clustering method applied on the Principal Components of the $\boldsymbol{U}$ matrix was agglomerative Hierarchical Clustering (HCPC), because this method does not depend on the prior selection of the number of the clusters, or a random initialization, as for example k-means does (Friedman *et al.*, 2001). HC requirements are the selection (a) of the dissimilarity measure, for which the Euclidean distance was used:

$$d(x_i, y_i) = \sqrt{\sum_{i=1}^{12} (x_i - y_i)^2} \tag{5}$$

where x_i and y_i are the dimensionless cumulative precipitation vectors of two different rainstorms; and (b) of the agglomeration method, where the Ward's minimum variance criterion was selected, an algorithm that minimizes the total within-cluster variance (Husson *et al.*, 2017), as implemented in the R language (Murtagh & Legendre, 2014).

At the beginning of the algorithm, the number of the clusters is equal to the number of data points (all clusters contain a single point). At every step, the algorithm finds the pair of clusters that result after merging to the minimum increase of the total within-cluster-variance, which is expressed as the sum of squared differences between the clusters' centers. Finally, all clusters are combined to one cluster that contains all the data using a hierarchical method.

The result from HCPC, a tree-based representation of the DCHs, was used to obtain the optimal number of clusters. At each step of the Algorithm the dendrogram is cut into different groups of DCHs and the center of each group represents a different distribution curve. These curves, for all

possible pairs, were tested whether they are drawn from the same distribution using the two-sample Kolmogorov-Smirnov test (William, 1971).

Because of the multiple pairwise tests, the p-values that resulted are adjusted using the Benjamini and Hochberg method, which controls the false discovery rate (Benjamini & Hochberg, 1995). If any of the produced hyetographs' p-values is not smaller than a predefined significance level α, the procedure stops and the optimal number of clusters is found.

2.3 Clustering tendency

Regarding the problem of classification of the DCHs, initially, and because all the clustering algorithms can return clusters, even if there is no structure in the used data, the Hopkins index, H, (Banerjee & Dave, 2004; Hopkins & Skellam, 1954) or clustering tendency was on $\boldsymbol{U}$. H can be used to test the null hypothesis of randomly and uniformly distributed data, generated by a Poisson point process and is calculated with:

$$H = \frac{\sum_{j=1}^{m} u_j^d}{\sum_{j=1}^{m} w_j^d + \sum_{j=1}^{m} u_j^d} \quad (6)$$

where when X is a collection of n data points that have d dimensions, a random sample from X without replacement with members x_i ($i = 1\ to\ m,\ m \ll n$) is formed and Y is a set of uniformly random data points, also with d dimensions and members y_j ($j = 1\ to\ m$), u_j in turn is the Euclidean distance from y_j to its nearest neighbor in X and w_j is also the Euclidean distance from x_i to its nearest neighbor in X. A value of H close to one, indicates that the data are highly clustered, 0.5 indicates randomly distributed data and zero indicates regularly spaced data (Theodoridis & Koutroumbas, 2009b).

3. Results and Discussion

3.1. Algorithm results

Using the calculated CD-values, that showed a temporal variation during summer months (Vantas *et al.*, 2018), a population of 1,622 out of 25,377 extracted rainstorms met the criteria of minimum duration and cumulative height. From PCA it is concluded that the use of the first two dimensions

explains 78.5% of total variance and the first 15 explains 99.5%. That means that it was possible to compress the DCHs' data and visualize them using only two dimensions, without losing much from their information. These facts are illustrated in Figure 3, which presents the resulting Scree Plot (Cattell, 1966) for the first ten dimensions of PCA.

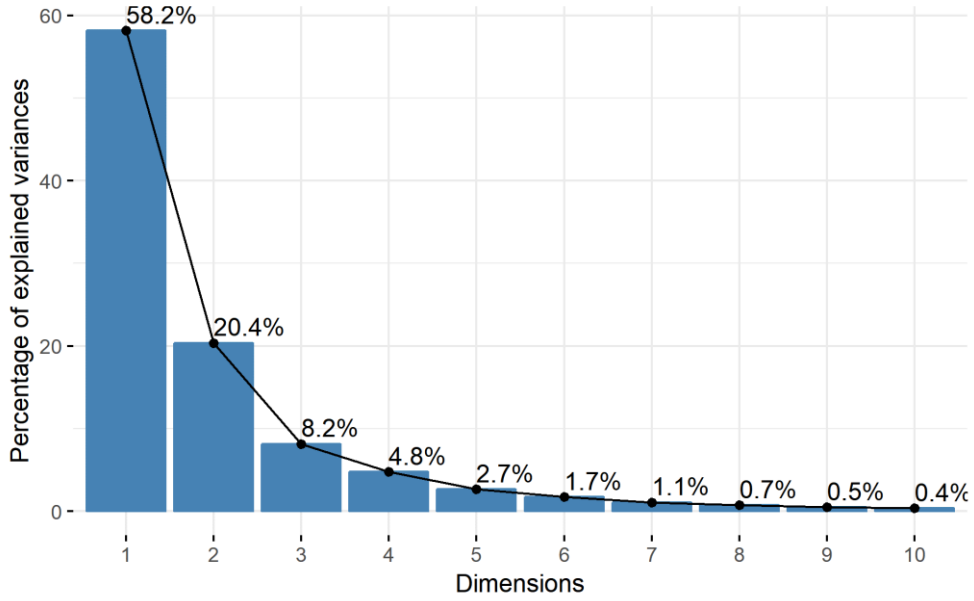


Figure 3. The Scree Plot using the first ten dimensions coming from applying PCA on the DCHs' data.

The application of the proposed Algorithm identified four clusters (Figure 4). Some of their statistics are presented in Table 2, and their monthly occurrence in Figure 5. The first cluster has notable higher average maximum 30 min duration's intensity and the highest variance in monthly occurrence. In Figure 5 the clusters' 10th, 50th and 90th percentiles are shown with the DCHs that belong to them. The percentiles' values of the clusters are given in Table 3 and Figure 6.

After developing distribution curves for each station and for every month, correlation matrices were computed, utilizing Pearson's r coefficient (Helsel & Hirsch, 2002), using the respective DCHs per cluster. These matrices showed very high similarity between a) the curves per station with $r \geq 0.98$, despite the missing values issues of the used dataset, and b) the curves per month with $r \geq 0.95$. On the basis of the above results, it may be concluded that these curves are representative for the given study area.

Table 2. Average values of occurrence of clusters, duration, precipitation height and maximum 30 min duration's intensity of clusters' rainstorms.

Cluster	Occurrence (%)	Duration (hr)	Prec. (mm)	$I30_{max}$ (mm/hr)
1	12.50	16.25	16.5	20.1
2	32.80	18.75	19.4	13.0
3	39.50	19.5	19.5	12.4
4	15.20	16.5	18.5	16.8

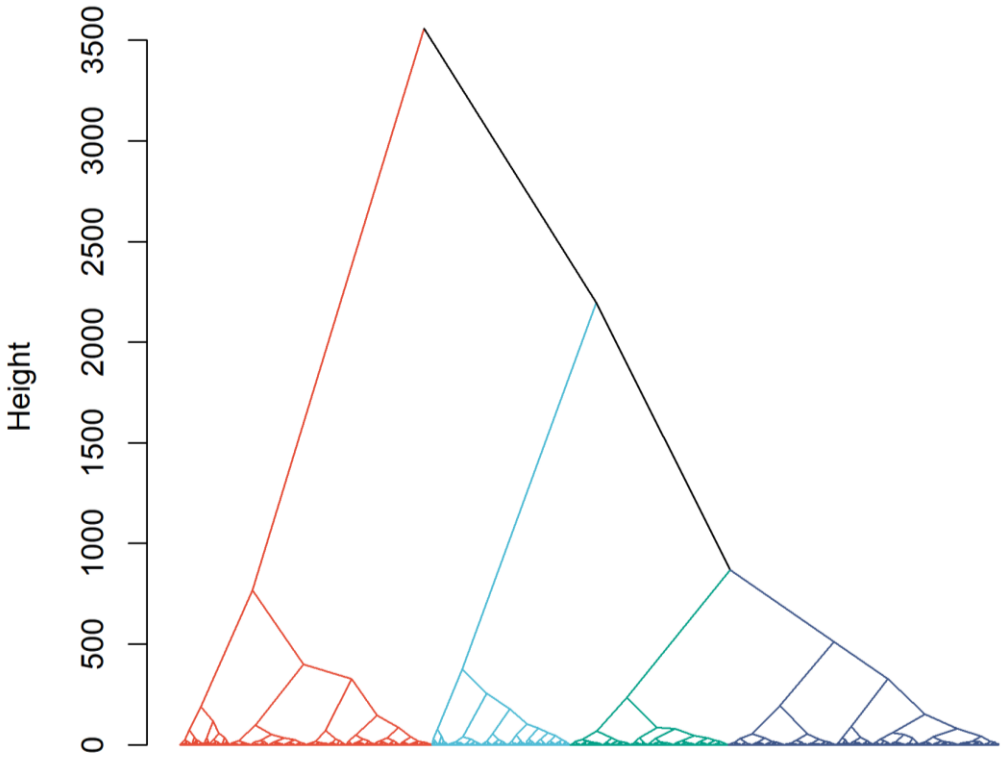


Figure 4. This dendrogram shows the tree-based representation of the DCHs that is produced from HCPC. Applying the Algorithm and moving down, the tree is cut into different clusters until the optimal number of them is found. The four optimal clusters are symbolized with different colors.

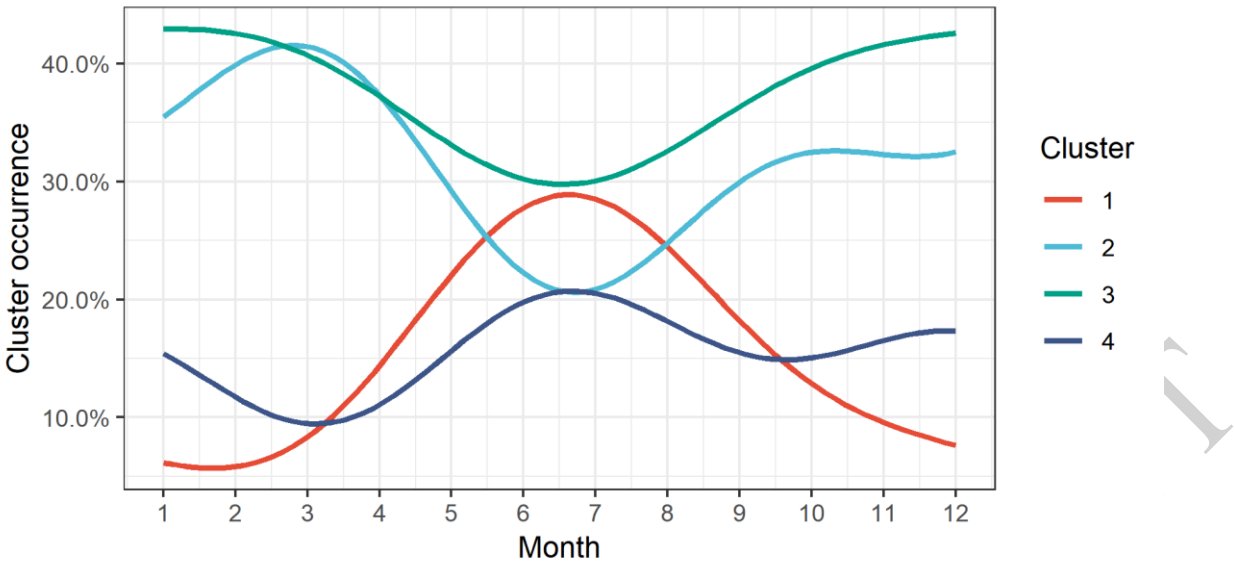


Figure 5. This plot presents the variability of clusters' monthly occurrence from HCPC.

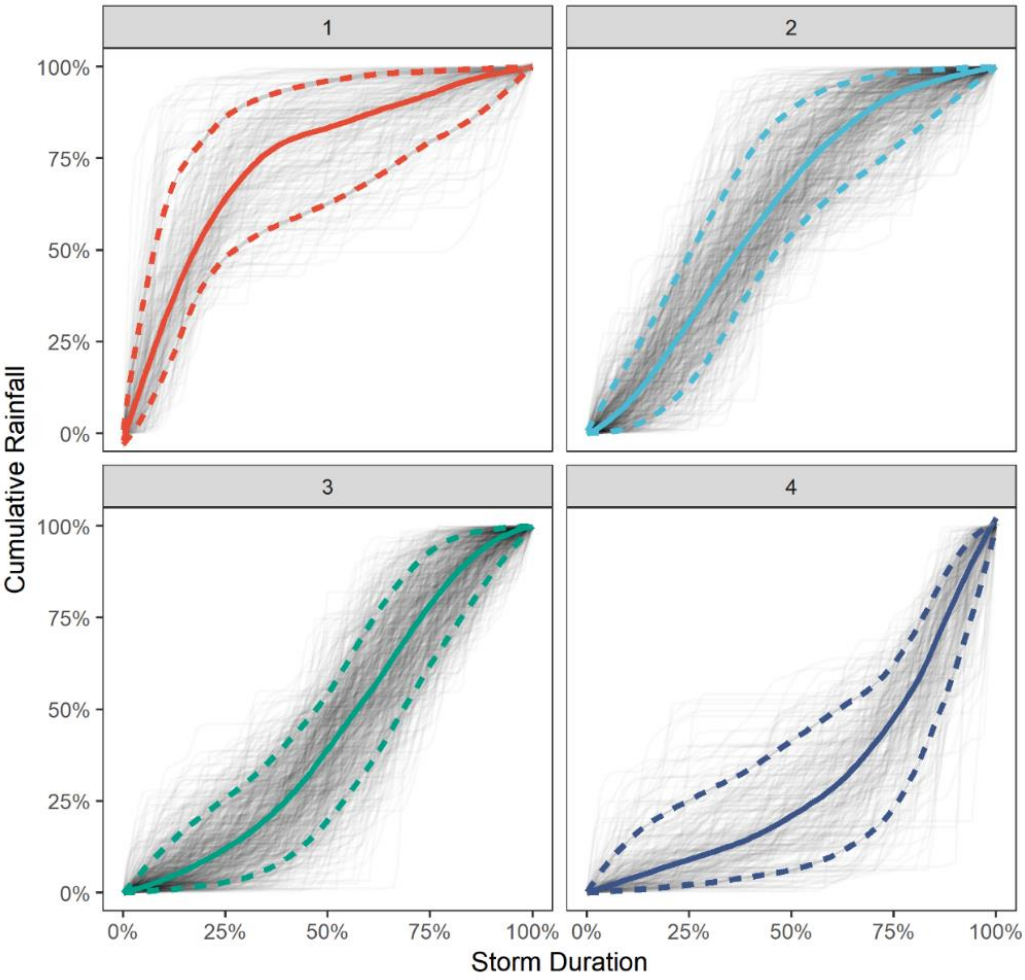


Figure 6. Results from the proposed Algorithm. With colors are presented the 10th, 50th (solid line) and 90th-percentiles dimensionless precipitation curves derived from the four optimal clusters. With grey lines are shown the DCHs of each cluster. The four panels depict the four different clusters.

Table 3. The percentiles' values of the DCHs as they are classified as clusters from HCPC. SD is an abbreviation for the dimensionless storm duration.

SD (%)	Cluster 1			Cluster 2			Cluster 3			Cluster 4		
	10th	50th	90th	10th	50th	90th	10th	50th	90th	10th	50th	90th
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5.0	3.9	15.7	38.4	0.4	3.1	10.0	0.4	1.7	6.6	0.4	1.9	8.1
10.0	15.0	31.4	63.5	1.6	8.1	19.2	0.8	3.6	12.4	0.8	3.7	14.3
15.0	29.8	44.9	73.0	3.9	14.9	28.5	1.3	6.0	17.3	1.1	5.5	18.7
20.0	41.8	55.7	80.4	8.0	22.4	38.2	2.0	8.9	21.8	1.6	7.4	22.2
25.0	47.8	64.5	86.8	13.5	30.5	48.6	2.9	12.0	26.0	2.2	9.2	25.3
30.0	52.0	71.4	89.6	20.7	38.9	58.6	4.2	15.7	30.2	2.7	10.9	28.2
35.0	55.8	76.7	92.2	29.9	46.7	67.6	6.2	20.1	35.1	3.5	12.7	31.0
40.0	57.5	80.0	93.7	39.5	54.5	76.9	9.6	25.5	41.1	4.3	14.9	33.8
45.0	60.3	81.8	95.2	47.5	61.8	83.5	13.8	31.6	47.3	5.2	17.6	37.7
50.0	62.7	83.5	96.2	54.0	68.7	88.2	19.6	39.2	54.2	6.1	21.1	41.7
55.0	65.2	85.3	96.9	60.0	75.1	92.2	26.9	47.1	63.4	7.7	24.4	45.0
60.0	68.7	87.4	97.8	64.7	80.4	94.5	34.4	54.2	73.3	10.1	28.3	49.4
65.0	71.8	89.1	98.3	69.0	84.9	96.0	43.4	62.4	80.9	13.0	33.6	52.6
70.0	76.0	90.8	98.5	72.8	89.1	97.4	53.1	71.0	87.8	16.9	40.3	56.4
75.0	79.5	92.5	98.7	77.6	92.3	98.3	62.2	78.5	93.3	22.8	47.7	62.3
80.0	82.0	94.8	99.1	82.1	94.3	98.9	70.7	85.4	96.3	32.4	56.1	70.8
85.0	85.5	96.6	99.3	86.7	96.0	99.3	79.0	91.3	98.1	46.2	67.0	81.3
90.0	89.5	97.7	99.5	91.2	97.5	99.5	86.9	95.5	99.1	60.2	81.1	91.7
95.0	94.6	98.8	99.7	95.6	98.8	99.8	94.0	98.3	99.6	78.0	92.7	98.3
100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

3.2. Clustering tendency and validation

The computed value of the Hopkins index *h* was 0.88, so the null hypothesis of random data was safely rejected. The previous result indicated that there was physical meaning in the categorization of rainstorms for the given dataset. As a comparison to selection of the optimal number of clusters of the proposed Algorithm, the 30 different indices that are used to determine the number of clusters in data, among others the popular "gap statistic" (Tibshirani *et al.*, 2001) and the "silhouette index" (Rousseeuw, 1987), listed in (Charrad *et al.*, 2014) were applied. Among all indices, six proposed two and another six proposed also four as the best number of clusters. Using the majority rule, and adding as a vote the results from the proposed Algorithm, the best number of clusters is four with seven votes (Figure 7). The previous result indicated that the proposed Algorithm identified as other widely used methods the optimal number of clusters, with the advantage that the centers of these clusters had statistically significant differences, a desirable feature of rainfall distribution curves.

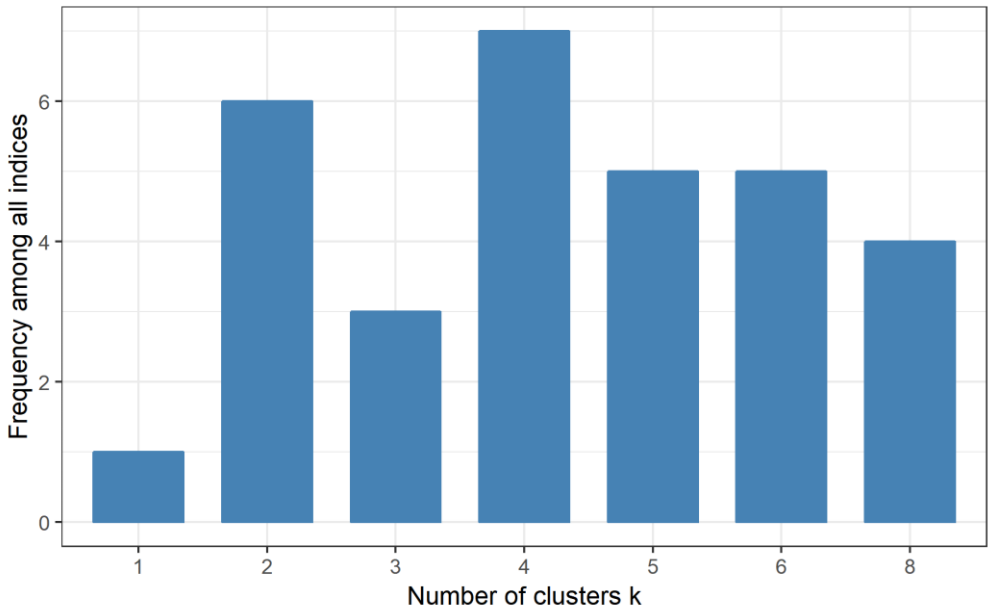
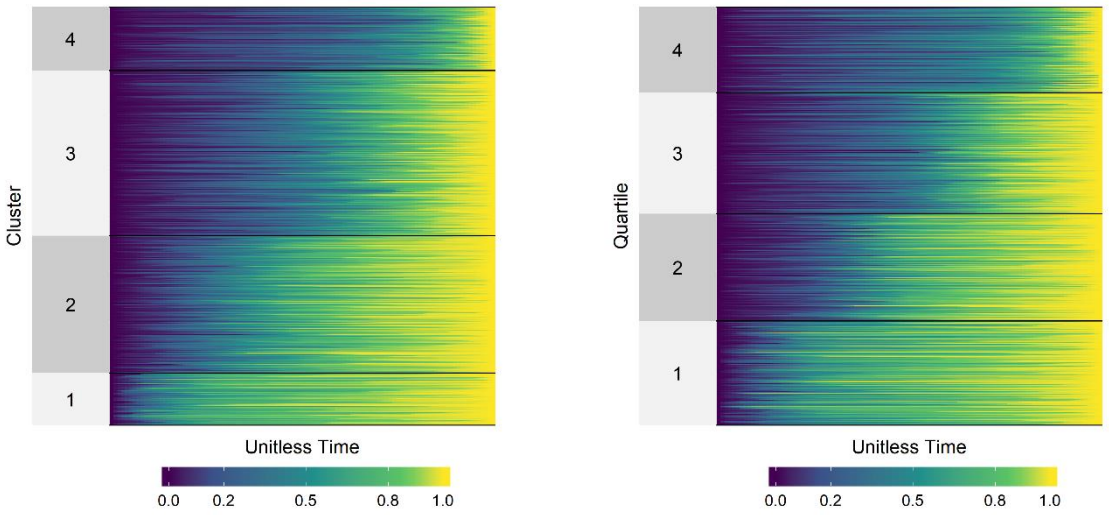


Figure 7. The above plot presents the frequency among all 30 indices used plus the results from the proposed Algorithm, for the determination of the optimal number of clusters.

3.2. Comparing Huff's curves with the proposed Algorithm's results

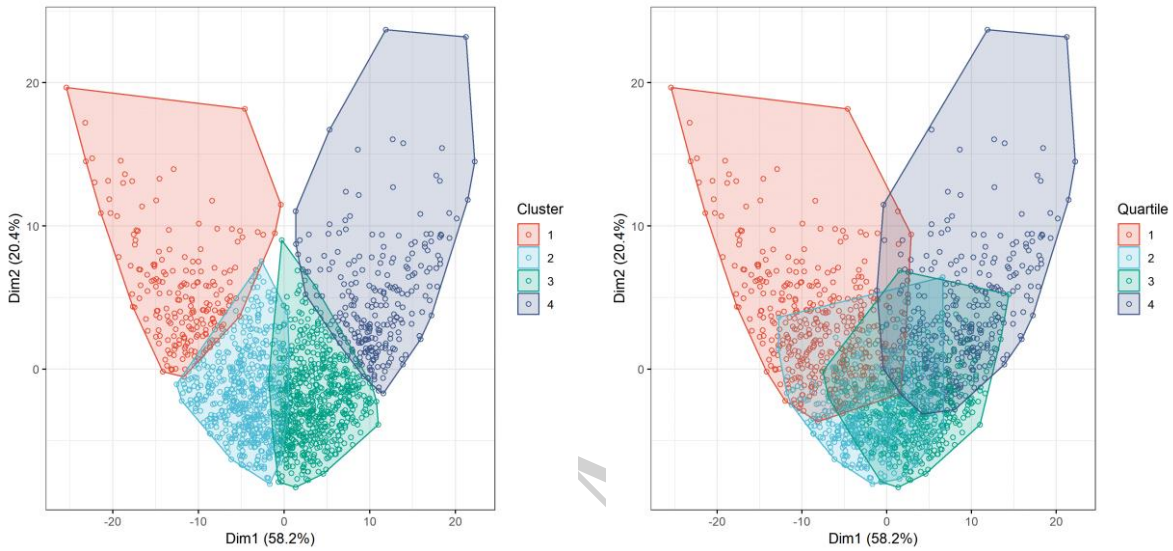
The internal structure of data can be seen in Figure 8 where the vectors of DCHs were re-ordered based on their cluster from the proposed Algorithm or their quartile that belong into. From this figure it can be seen that the two different methods produced different results.



(a) (b)

Figure 8. The above two diagrams present the different ways of classification of DCHs from HCPC (a) and Huff's classification (b). In both plots the color range represents the value of dimensionless cumulative rainfall and the x axis is the dimensionless time.

Given that the first two dimensions from PCA can be used to present DCHs as points, in Figure 9 the DCHs are presented with their corresponding ellipses around them as a method of visual internal validation of the clustering results. These ellipses create the areas of the clusters formed both from HCPC, which separates the points clearly, and Huff's classification, where the ellipses are overlapping.



(a) (b)

Figure 9. In these diagrams DCH's are plotted using the first two principal components and ellipses around the clusters. The results from HCPC are given in (a) and from Huff's classification in (b).

Table 4. Adjusted p-values test using the Benjamini & Hochberg method coming from the two-sample Kolmogorov-Smirnov tests. In (a) are tested the design curves from the HCPC and in (b) from Huff's classification. C is an abbreviation for Cluster and Q for Quartile.

	C 2	**C 3**	**C 4**
C 1	0.01	$4 \cdot 10^{-7}$	$2 \cdot 10^{-14}$
C 2		0.04	10^{-6}
C 3			0.04

(a)

	Q 2	**Q 3**	**Q 4**
Q 1	0.10	$4 \cdot 10^{-5}$	$3 \cdot 10^{-10}$
Q 2		0.11	$4 \cdot 10^{-5}$
Q 3			0.12

(b)

Also, all possible pairs of curves, were tested whether they drawn from the same distribution using the two-sample Kolmogorov-Smirnov test (their adjusted p-values using the Benjamini and Hochberg method). These statistical tests were used as a relative validation method of the clustering results. Three pairs of the Huff's curves failed to reject the hypothesis that are drawn from the same distribution for both $\alpha = 0.05$ and $\alpha = 0.10$ in contrast to HCPC results (Table 4). In other words, for

291 the given dataset, Huff's classification failed to produce statistical independent distribution curves
292 in contrast to the ones coming from the proposed Algorithm.

293 **3. Conclusions**

294 A novel method is presented in order to classify statistically independent heavy rainstorm events
295 and create a limited number of intra-storm temporal distribution curves, a fact that is interesting
296 both from the methodology point of view and in regard to practical applications. More specifically,
297 the conclusions from the followed analytical steps can be briefly outlined as follows. Principal
298 Components Analysis showed that it is possible to compress the hyetograph data to a few
299 dimensions without losing much information and consequently to apply the proposed method to big
300 data-sets. Clustering tendency analysis showed that the DCHs data set used with the proposed
301 Algorithm contains meaningful clusters (i.e. non-random structures). The relative clustering
302 validation analysis showed that the proposed method, in a majority voting scheme with another 30
303 indices, produced the optimal number of clusters. Internal structure validation and statistical testing
304 showed that the proposed Algorithm provides better classification of the DCHs than the established
305 Huff's quartile classification. Four representative distribution curves were produced and such
306 hyetographs have not been derived in Greece so far, especially in a way that covers the various
307 Water Divisions. The production of only a limited number of representative distribution curves
308 offers considerable advantages for practical purposes and the method presented here replaces more
309 traditional methods, because it is more rational, as it is fully unsupervised and it requires no prior
310 empirical knowledge. This is achieved, because with the proposed Algorithm no human
311 intervention or bias is involved in the selection of the clusters.

312 **Acknowledgments:**

313 The analysis and the algorithm were implemented in the R language (R Core Team, 2019), using
314 the packages: hydroscoper (Vantas, 2018a), hyetor (Vantas, 2018b), nbclust (Charrad et al., 2014)
315 FactoMineR (Lê et al., 2008), and factoextra (Kassambara & Mundt, 2017).

316

References

Abu-Mostafa, Y. S., Magdon-Ismail, M. & Lin, H.-T. (2012) *Learning from data*, Vol. 4. AMLBook Singapore.

Azli, M. & Rao, A. R. (2010) Development of Huff curves for Peninsular Malaysia. *Journal of Hydrology* **388**(1–2), 77–84. doi:10.1016/j.jhydrol.2010.04.030

Banerjee, A. & Dave, R. N. (2004) Validating clusters using the Hopkins statistic. *2004 IEEE International Conference on Fuzzy Systems (IEEE Cat. No.04CH37542)*. IEEE. doi:10.1109/fuzzy.2004.1375706

Benjamini, Y. & Hochberg, Y. (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**, 289–300.

Bezak, N., Šraj, M., Rusjan, S. & Mikoš, M. (2018) Impact of the rainfall duration and temporal rainfall distribution defined using the Huff curves on the hydraulic flood modelling results. *Geosciences* **8**(2), 69.

Bonta, J. (2001) Characterizing and estimating spatial and temporal variability of times between storms. *Transactions of the ASAE* **44**(6), 1593.

Bonta, J. (2004) Development and utility of Huff curves for disaggregating precipitation amounts. *Applied engineering in agriculture* **20**(5), 641.

Bonta, J. (2004) Stochastic simulation of storm occurrence, depth, duration, and within-storm intensities. *Transactions of the ASAE* **47**(5), 1573.

Bonta, J. & Rao, A. (1987) Factors affecting development of Huff curves. *Transactions of the ASAE* **30**(6), 1689–1693.

Bonta, J. V. & Shahalam, A. (2003) Cumulative storm rainfall distributions: comparison of Huff curves. *Journal of Hydrology (New Zealand)* **42**(1), 65–74.

Cattell, R. B. (1966) The scree test for the number of factors. *Multivariate behavioral research* **1**(2), 245–276.

Charrad, M., Ghazzali, N., Boiteau, V. & Niknafs, A. (2014) NbClust : An R Package for Determining the Relevant Number of Clusters in a Data Set. *Journal of Statistical Software* **61**(6). doi:10.18637/jss.v061.i06

Dolšak, D., Bezak, N. & Šraj, M. (2016) Temporal characteristics of rainfall events under three climate types in Slovenia. *Journal of Hydrology* **541**, 1395–1405. doi:10.1016/j.jhydrol.2016.08.047

Feng, Y. & Hamerly, G. (2007) PG-means: learning the number of clusters in data. *Advances in neural information processing systems*, 393–400.

Friedman, J., Hastie, T. & Tibshirani, R. (2001) *The elements of statistical learning*, Vol. 1. Springer series in statistics New York.

Guo, X. B., Wang, Z. Q. & Zhang, R. L. (2001) Study on temporal distribution of rainfall erosivity and daily rainfall erosivity model in red soil area of Zhejiang. *Journal of Soil and Water Conservation* **15**(3), 35–37.

Hartigan, J. A. & Wong, M. A. (1979) Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **28**(1), 100–108.

Helsel, D. R. & Hirsch, R. M. (2002) Statistical Methods in Water Resources. In: *Hydrologic Analysis and Interpretation* Techniques of Water-Resources Investigations of the United States Geological Survey. U.S. Geological Survey.

Hopkins, B. & Skellam, J. G. (1954) A new method for determining the type of distribution of plant individuals. *Annals of Botany* **18**(2), 213–227.

Hsu, K.-C. & Li, S.-T. (2010) Clustering spatial–temporal precipitation data using wavelet transform and self-organizing map neural network. *Advances in Water Resources* **33**(2), 190–200.

Huff, F. A. (1967) Time distribution of rainfall in heavy storms. *Water Resources Research* **3**(4), 1007–1019. doi:10.1029/WR003i004p01007

Husson, F., Lê, S. & Pagès, J. (2017) *Exploratory Multivariate Analysis by Example Using R*, 2nd ed. New York, NY, USA: Chapman and Hall/CRC.

Jiang, P., Yu, Z., Gautam, M. R., Yuan, F. & Acharya, K. (2016) Changes of storm properties in the United States: Observations and multimodel ensemble projections. *Global and Planetary Change* **142**, 41–52.

Jolliffe, I. T. (1986) Principal Component Analysis and Factor Analysis. In: *Principal Component Analysis*, 115–128. Springer.

Kassambara, A. & Mundt, F. (2017) *factoextra: Extract and Visualize the Results of Multivariate Data Analyses*.

Koutsoyiannis, D. (1994) A stochastic disaggregation method for design storm and flood synthesis. *Journal of Hydrology* **156**(1–4), 193–225. doi:10.1016/0022-1694(94)90078-7

Lê, S., Josse, J. & Husson, F. (2008) FactoMineR: A Package for Multivariate Analysis. *Journal of Statistical Software* **25**(1), 1–18.

Lin, G.-F. & Wu, M.-C. (2007) A SOM-based approach to estimating design hyetographs of ungauged sites. *Journal of Hydrology* **339**(3–4), 216–226.

Loukas, A. & Quick, M. C. (1996) Spatial and temporal distribution of storm precipitation in southwestern British Columbia. *Journal of hydrology* **174**(1–2), 37–56.

MacQueen, J. (1967) Some methods for classification and analysis of multivariate observations. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Vol. 1, 281–297. Oakland, CA, USA.

Milligan, G. W. & Cooper, M. C. (1985) An examination of procedures for determining the number of clusters in a data set. *Psychometrika* **50**(2), 159–179.

Ministry of Environment and Energy. (2013) Management plan of Thracian Water Division. Ministry of Environment and Energy.

Murtagh, F. & Legendre, P. (2014) Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? *Journal of classification* **31**(3), 274–295.

Nojumuddin, N. S. & Yusop, Z. (2015) Identification of Rainfall Patterns in Johor. *Applied Mathematical Sciences* **9**(38), 1869–1888.

Pan, C., Wang, X., Liu, L., Huang, H. & Wang, D. (2017) Improvement to the Huff Curve for Design Storms and Urban Flooding Simulations in Guangzhou, China. *Water* **9**(6), 411. doi:10.3390/w9060411

Pearson, K. (1901) LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **2**(11), 559–572.

Perica, S., Kane, D., Dietz, S., Maitaria, K., Martin, D., Pavlovic, S., Roy, I., et al. (2012) Precipitation-Frequency Atlas of the United States 128.

R Core Team. (2019) *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Restrepo-Posada, P. J. & Eagleson, P. S. (1982) Identification of independent rainstorms. *Journal of Hydrology* **55**(1–4), 303–319.

Rousseeuw, P. J. (1987) Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* **20**, 53–65.

Sheikholeslami, G., Chatterjee, S. & Zhang, A. (1998) Wavecluster: A multi-resolution clustering approach for very large spatial databases. *VLDB*, Vol. 98, 428–439.

Theodoridis, S. & Koutroumbas, K. (2009) *Pattern recognition*, 4. ed. Amsterdam: Elsevier Acad. Press.

Theodoridis, S. & Koutroumbas, K. (2009) *Pattern Recognition*. Burlington,MA,USA: Academic Press.

Tibshirani, R., Walther, G. & Hastie, T. (2001) Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **63**(2), 411–423. doi:10.1111/1467-9868.00293

Vantas, K. (2018) hydroscoper: R interface to the Greek National Data Bank for Hydrological and Meteorological Information. *Journal of Open Source Software* **3**(23), 625. doi:10.21105/joss.00625

Vantas, K. (2018) hyetor: R package to analyze fixed interval precipitation time series. doi:10.5281/zenodo.1403156

Vantas, K. & Sidiropoulos, E. (2017) Imputation of erosivity values under incomplete rainfall data by machine learning methods. *European Water* **57**, 193–197.

Vantas, K., Sidiropoulos, E. & Loukas, A. (2019) Robustness Spatiotemporal Clustering and Trend Detection of Rainfall Erosivity Density in Greece. *Water* **11**(5), 1050. doi:10.3390/w11051050

Vantas, K., Sidiropoulos, E. & Vafeiadis, M. (2019) A data driven approach for the temporal classification of heavy rainfall using Self-Organizing Maps. *EGU General Assembly 2019*, Vol. 21, 1. Vienna, Austria.

Vantas, K., Sidiropoulos, E. & Vafiadis, M. (2018) Rainfall Temporal Distribution in Thrace by Means of an Unsupervised Machine Learning Method. *Protection and Restoration of the Environment XIV*, 555–564.

Veneziano, D. & Villani, P. (1999) Best linear unbiased design hyetograph. *Water Resources Research* **35**(9), 2725–2738.

Ward, J. (1963) Hierarchical grouping to optimize an objective function. *Journal of the American statistical association* **58**(301), 236–244.

William, J. C. (1971) *Practical nonparametric statistics*. John Wiley & Sons, New York.

Williams-Sether, T., Asquith, W. H., Thompson, D. B., Cleveland, T. G. & Fang, X. (2004) Empirical, dimensionless, cumulative-rainfall hyetographs developed from 1959-86 storm data for selected small watersheds in Texas. *Statistical Characteristics of Storm Interevent Time, Depth, and Duration for Eastern New Mexico, Oklahoma, and Texas*.

Yin, S., Xie, Y., Nearing, M. A., Guo, W. & Zhu, Z. (2016) Intra-storm temporal patterns of rainfall in China using Huff curves. *Transactions of the ASABE* **59**(6), 1619–1632.

Zambelli, A. (2016) A Data-Driven Approach to Estimating the Number of Clusters in Hierarchical Clustering. *arXiv:1608.04700 [cs, q-bio, stat]*. Retrieved from http://arxiv.org/abs/1608.04700

Zeimetz, F., Schaefli, B., Artigue, G., Hernández, J. G. & Schleiss, A. J. (2018) Swiss Rainfall Mass Curves and their Influence on Extreme Flood Simulation. *Water Resour Manage* **32**(8), 2625–2638. doi:10.1007/s11269-018-1948-y

Zhou, S., Xu, Z. & Liu, F. (2017) Method for Determining the Optimal Number of Clusters Based on Agglomerative Hierarchical Clustering. *IEEE Transactions on Neural Networks and Learning Systems* **28**(12), 3007–3017. doi:10.1109/TNNLS.2016.2608001