European Journal of Geography Volume 10, Number 2:27-41, June 2019

K-MEANS CLUSTERING OF A SOIL SAMPLING SCHEME WITH DATA ON THE MORPHOGRAPHY OF THE OGOSTA VALLEY NORTHWESTERN BULGARIA

Assen TCHORBADJIEFF
Bulgarian Academy of Science, Institute of Mathematics and Informatics, Sofia, Bulgaria
atchorbadjieff@math.bas.bg

Tsvetan KOTSEV
Bulgarian Academy of Science, National Institute of Geophysics, Geodesy and Geography, Sofia, Bulgaria
tsvetankotsev@mail.bg

Velimira STOYANOVA
Bulgarian Academy of Science, National Institute of Geophysics, Geodesy and Geography, Sofia, Bulgaria
stoyanovavelimira@gmail.com

Emilia TCHERKEZOVA
Bulgarian Academy of Science, National Institute of Geophysics, Geodesy and Geography, Sofia, Bulgaria
et@geophys.bas.bg

Abstract

The spatial distribution of 665 soil sampling sites in the arsenic contaminated floodplain of the Ogosta River in the Northwest of Bulgaria is analysed against geomorphological parameters computed from a precise digital terrain model. The study aims at partitioning and classifications of hidden patterns of the morphographic features of the river floodplain, which to be used for the explanation of the arsenic dispersal in the polluted soils at a further stage. The field sites are split into 4 clusters using K-means algorithm with the following variables: elevation, distance to the river, vertical distance to channel network, multiresolution index of valley bottom flatness and a modified topographic SAGA wetness index. It is found that each cluster is related to a distinct area in the valley and is in good agreement with the distribution of the previously determined geomorphological units, as well as with the extent of a simulated historic flood.

Keywords: *Spatial clustering, K-means clustering, River pollution, Digital Terrain Model (DTM), Floodplain, Ogosta River*

1. INTRODUCTION

The spatial distribution of trace metals in contaminated soil of river valleys is often determined by pollutant transport with suspended river sediment during flood events. This is the usual case when rivers are affected by mining activities due to continuous or accidental release of mine waste (Bird et al., 2010). Once released, the metal contaminants are mainly transported in a particulate-associate form (Martin and Meybeck, 1979). Their dispersal, storage and remobilization in the fluvial system can be directly related to sediment transport processes, styles of the river channel and floodplain sedimentation, and flooding regime (Macklin et al., 2006). The relationship between the distribution of heavy metals in the soils and the fluvial forms of relief in polluted floodplains is presented in numerous researches (Taylor and Hudson-

Edwards, 2007; Clement et. al., 2017). This relationship is the basis of the geomorphological-geochemical approach for exploring the distribution of heavy metals in river valleys, developed by Macklin et al. (2006). It takes into consideration parameters related to sediment transport and accumulation such as height above the river channel, distance from the river bank, frequency of flooding, sediment age and sedimentation conditions. Some studies on metal contamination of river systems link pollutant dispersal with topographic features of the river floodplain (Dennis et al., 2003; Ciszewski et al., 2012), while other researches consider mostly the extent of floods (Brewer et al., 2005).

Usually, similar researches in geomorphology and Earth surface dynamics rely on advanced computational methods for analysis of different types of geomorphological and geomorphographic units, alone or in combination with other geomorphometric parameters. For instance, a Multiresolution Index of Valley Bottom Flatness (MRVBF) is used to delineate geomorphologic and hydrologic units, and for mapping depositional areas (Gallant and Dowling, 2003). A simple slope-discharge model using Support Vector Machine (SVM) is introduced for mapping and modelling of the channel patterns of the Columbia River basin, USA (Beechie and Imaki, 2014). A nested hierarchical scheme is used for characterization several coastal river systems in New South Wales (Brierley and Fryirs, 2000). The models show river interpretation of character and behaviour with inter-related scales based on catchments, landscape units, river styles and geomorphic units. Similarly, a hierarchical scheme is used for estimation of habitat dependence on river geomorphology (Thomson et al., 2001). Other interesting results are obtained after comparison of K-means clustering with hierarchical one for 3D data from Terrestrial Laser Scanning of Landslide in Dunning et al. (2009).

The selected methods for data analysis are case dependent mainly due to the specificity of data and uniqueness of observed river beds and riparian terrain. Moreover, the usage of more sophisticated methods has become a very common part of similar research due to the rapid development of Machine Learning (ML) methods and growing data volumes. Particularly, in geomorphology, three different approaches are mainly used for systematic analysis of available data complexity - classification and cataloguing; cluster analysis; and regression and interpolation (Valentine and Kalnins, 2016). The classification is usually selected for searching any predefined patterns, as demonstrated by the channel modelling for the Columbia River basin, USA (Beechie and Imaki, 2014). Conversely, clustering is considered when the process is without knowledge of clusters and their number, i.e. it is unsupervised. The output is data partition in two or more parts with similar patterns, but with dimensions reduction. Finally, regression and interpolation methods are used to estimate relations between physical parameters and to produce predictions.

In more complex cases, regression could be applied using the outcome of data classification or clustering. This is the case for the pollutant data from the mining-affected valley of the Ogosta River in the north-western part of Bulgaria, obtained in several campaigns in one decade. The collected data increased its volume over time and spread over a large non-homogeneous region. In this regard, the present study aims at dividing the valley floor into more homogeneous sections regarding the conditions for metal-contaminant dispersal. For this purpose, the K-means clustering algorithm is selected for unsupervised grouping of the available observed sites as a required step before implementation of regression to reveal the spatial distribution of heavy metals in soil.

2. MATERIALS AND METHODS

2.1. Study area

The investigated area is located in the upper stretch of the Ogosta River valley between the village of Martinovo and the Ogosta dam lake (Figure 1). It covers an area of the valley floor of around 14.5 sq. km with a mean elevation of 491 m and an average inclination of slopes about 3˚.

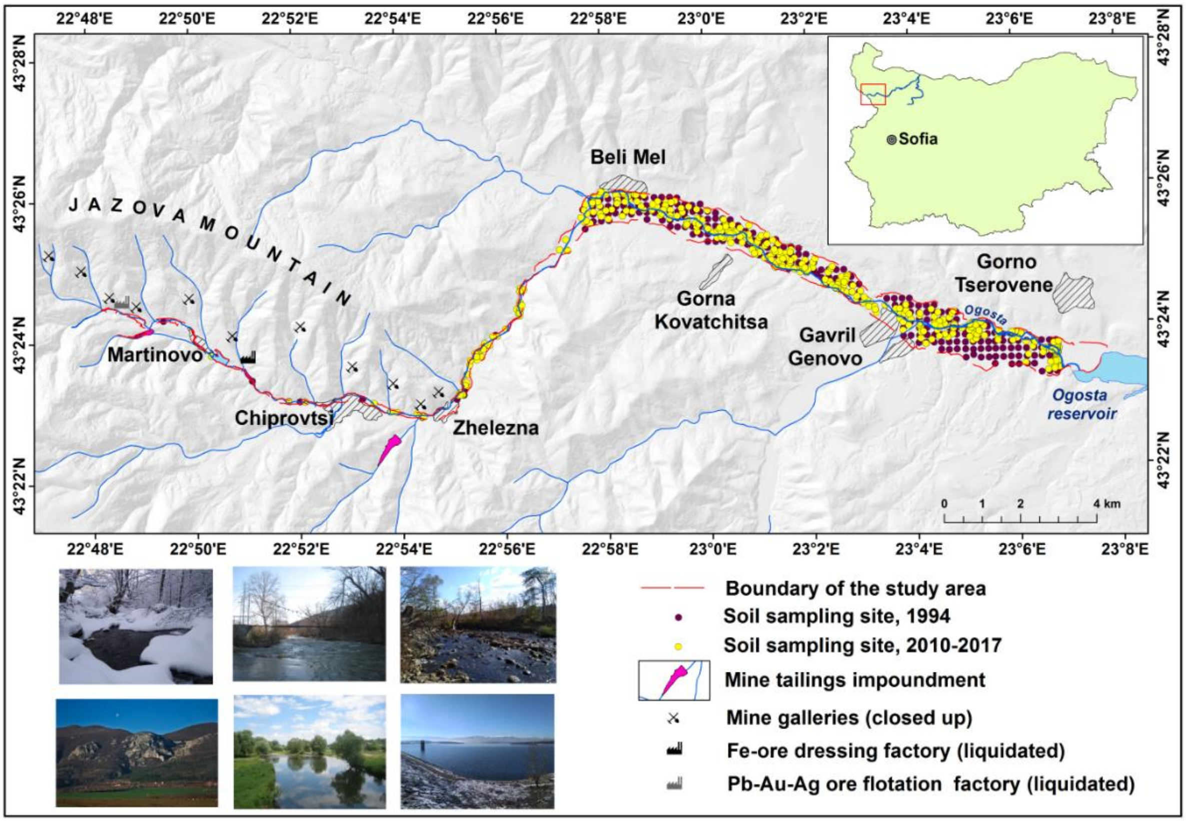

Figure 1. Designation of the mines and the soil sampling sites within the investigated section of the Ogosta Valley

The valley extends over a part of the Western Balkan mountain range and Western Fore-Balkan. It includes territories of three municipalities: Chiprovtsi, Georgi Damyanovo and Montana. Extraction and dressing of iron-ore and lead-silver-ore took place near the town of Chiprovtsi in the upper reach of the Ogosta River from 1951 to 1999. Due to a tailings dam failure in 1964 and to the mine waste discharge into the Ogosta River in the period 1964-1979, the floodplain soil in the Ogosta Valley received significant amounts of arsenic and heavy metals (Jordanova et al., 2013).

2.2. Soil sampling network

This study uses the data from two investigations on soil contamination in the Ogosta Valley, conducted in 1994 (Spectroteh, 1994) and 2010-2017 (Mandaliev et al., 2014; Simmler et al., 2016). They are considered compatible due to the same area of investigation, and the comparable concentrations of trace elements found at adjacent soil sampling points of the two studies. The first research applies a regular soil sampling grid with cell size 200 x 200 m. It well characterizes the distribution of arsenic in the less polluted sections of the valley which are more distant from the Ogosta River. However, the sampling grid is not enough detailed to reveal the diverse spatial pattern of the intensive contamination in the lower floodplain next to the river. The sampling concept of the later investigation complements the field site network of the earlier study, focusing on the more contaminated areas. It takes into account primarily the floodplain morphography and does not follow a regular grid. The combined soil sampling network from the two studies has a higher density in the low river floodplain and is less detailed in the periphery of the bottom of the valley. Its irregular pattern reflects in a better way the controls of the contaminant dispersal in the valley, compared to the sampling nets of each of the two studies alone. After application of some data quality procedures, the final number of sampling locations is reduced from 699 to 665, with 254 observed sited during the campaign in 1994 and 411 points from this in 2012-2017.

2.3. Acquired data

A set of five geomorphometric variables are used to identify the sites from the combined soil sampling network (Table 1). The parameters were computed from a Digital Terrain Model (DTM) with pixel size 1x1 m, generated from Airborne Laser Scanning Data obtained in 2013. Some of them give information on the lateral and vertical distance from the Ogosta River as the primary source of soil contamination. Due to this, the model is centred geospatially around the river, and every site is described uniquely in all three coordinate directions. Other parameters characterize the potential of deposition of sediment and particulate matter within different fluvial landforms during inundation. Two additional parameters are applied to control the results of the clustering - delineated morphographic units of the valley floor and numerical estimates of the flooding in 1964.

Table 1. Computed geomorphological parameters and units.

Type	Parameter	Additional description
Variables used for clustering		
Numerical	Digital Terrain Model (DTM) (1x1 m)	Altitude above the sea level, in meters
Numerical	Distance to the river	The distance between a sampling site and the Ogosta River in the Cartesian coordinate system, in meters.
Numerical	Vertical distance to channel network (VDCN) *	Altitude above the channel network level, in meters.
Numerical index	Multiresolution index of valley bottom flatness (MRVBF) *	The index was computed from DTM (1x1 m) is used for mapping depositional areas (Gallant and Dowling, 2003).
Numerical index	Modified Topographic SAGA Wetness Index (mTWI)*	The index gives information about the potential soil moisture (Conrad et al., 2015).
Variables used for validation of the results of the clustering		
Numerical	Inundation depth	Simulated inundation depth during the flood in April 1964.
Category	Geomorphographic units (GMU)*	Identified floodplain units using cross-classification and tabulation of terrain classification index of lowlands (TCI_{low}) (Bock et al., 2007) and Vertical distance to channel network (in m) (Tcherkezova, 2015) using SAGA GIS (Conrad et al., 2015).

* Computed with SAGA GIS software (Conrad et al., 2015, www.saga-gis.org/)

Simulations of major historical flood events during the mining period are performed to outline the areas where the arsenic-contaminated river sediments are likely to have been accumulated. The inundation in April 1964 had the highest impact due to the entering in the river of 100 000 m^3 slurry with a high concentration of arsenic. Its span and water depth are used to characterize the spatial distribution of the contaminant in the floodplain soil. The hydrological modelling of the high flow event is performed with the software AGWA (Automated Geospatial Watershed Assessment) and its constituent model SWAT (Soil and Water Assessment Tool) (Arnold et al., 2012). The calibration of the model is conducted with daily precipitation data and river discharge records for several storm events in the Ogosta Valley in the period 2002 – 2005. The software HEC-RAS version 4.1.0 is applied for the hydraulic modelling of the historical flood events (Brunner, 2010). The model is calibrated with data on the extent of the river inundation on April 19, 2014.

Due to requirements of the K-means clustering algorithm, the selected model parameters for computations are constrained to combinations of only four numeric geomorphological variables and the distance to the river (see Table 1). The remaining two parameters, flooding estimates and the geomorphic units (GMU), are preserved for verification of differences between clusters. Because the GMUs are categorical data, they are excluded from direct computations (see Table 2). However, the GMUs represent very well the geomorphographic landform patterns and therefore their distribution in obtained clusters is used for validation. In addition, the selected data is restricted only to these sites which GMU units do not exceed the relative height above the channel network level up to 6,5–7 m. The estimates from flooding simulations are another useful control parameter. The dynamics of flooding is dependent on a number of parameters such as rainfall intensity, floodplain and channel morphology and deposits, river bank characteristics, geometry and dynamics of stream channels, and others. Thus, it is expected that geomorphologic units in cluster distribution are in good agreement with flooding predictions. However, because the amplitude of flooding is also dependent on the available water amount, the expectations of complete overlapping could be misleading.

Table 2. Description of GMU categorical values

GMU index	Description
100	Bankfull channel zone: stream bed, backwater areas, abandoned channels and valleys, 0 - 0,5 m
101	Bankfull channel zone, active floodplain (T_{0-l}) fragments with backwater areas, abandoned channels, anthropogenic areas near Ogosta reservoir, 0,5 – 1 m
102	Active floodplain (T_{0-l}) fragments with small locale depressions, abandoned channels, river banks of recent and abandoned channels, embankments (natural and anthropogenic), sand bars, levees, anthropogenic areas near Ogosta reservoir, 1–1,5 m
103	Active floodplain (T_{0-l}) fragments with small locale depressions, more or less coarse structure with convex micro-forms, sand bars, levees, low embankments (natural and anthropogenic), anthropogenic areas, 1,5–2 m
104	Active floodplain (T_{0-l}) fragments, more or less coarse structure with convex micro-forms, sand bars, leaves, low escarpments, embankments (natural and anthropogenic), bank of roads, anthropogenic areas, 2–2,5m
105	Active floodplain (T_{0-l}) with small convex micro-forms, low escarpments, river banks, embankments (natural and anthropogenic), anthropogenic areas, 2,5–3 m
200	Floodplain (T_{0-l}) (fragments) with coarse structure, escarpments, bank of roads, embankments (natural and anthropogenic), 3 – 3,5 m
201	Floodplain T_{0-h} (fragments) with coarse structure, alluvial fan deposits, escarpments, bank of roads, embankments (natural and anthropogenic), 3,5–4 m
202	Floodplain T_{0-h} (fragments) with coarse structure, alluvial fan deposits, escarpments, bank of roads, embankments (natural and anthropogenic), 4–4,5 m
203	Floodplain T_{0-h} (fragments) with coarse structure, alluvial fan deposits, escarpments, a bank of roads, embankments (natural and anthropogenic), 4,5–5 m
204	Floodplain T_{0-h} (fragments) with coarse structure, alluvial fan deposits, escarpments, bank of roads, embankments (natural and anthropogenic), 5–5,5 m

205	Floodplain $T_{0\text{-}h}$ (fragments) with coarse structure, alluvial fan deposits, escarpments, bank of roads, embankments (natural and anthropogenic), 5,5–6 m
206	Floodplain $T_{0\text{-}h}$ (fragments) with coarse structure, alluvial fan deposits, escarpments, bank of roads, embankments (natural and anthropogenic), 6–6,5 m
300	Escarpments, alluvial fan deposits, colluvial deposits, slopes, up to 6,5–7 m relative height

2.4. Methods

Dealing with similar large parameterized data is not trivial and classification of area stretched over large irregular grids is required as a first step in the analysis. Moreover, the data can be acquired from different sources, even from unexpected historic ones (Weis and Hrončešk, 2017). The expected result is a reduction of data dimensions by grouping of the observed sites by selected morphographic parameters. The K-means clustering is selected as the most straightforward and convenient solution in this case. An important property is that it converges very quickly and it is very effective with big data. This is a serious advantage compared to hierarchical clustering methods in cases of large multivariate datasets. Nevertheless, it is a method based on grouping data around K-number centres named as centroids. This decentralized structure enables concurrency between them and algorithm works by having the clusters compete with each other for the right to own the data points. This classifies the model as a competitive learning algorithm and any predefined classification is not required (MacKay, 2003).

The K-means clustering is an algorithm that split N-number of data points of an I-dimensional space (number of parameters) to k clusters. Every cluster consists of vectors x and any of them has I-number of elements x_i. The elements that constitute every cluster are centred around the centroid, parameterized with i-sized vector $m^{(k)}$ of mean values. The optimal distribution of the group of n points $\{x^{(n)}\}_{n\leq N}$ around the centroids is determined by the minimal distance between them and mean values. There are many different metrics for measuring this distance, mainly related to inter-data dependency. When observed data is independently distributed over irregular terrain, the Euclidean distance metric is good enough for designed computations. It is a quadratic distance between two objects *i* and *j*:

$$d_{i,j} = \sqrt{\frac{\sum_{k=1}^{I}\left(x_{ik} - x_{jk}\right)^2}{I}}$$

For initialization of the K-means algorithm, the number of expected mean values $m^{(k)}$ is required. The expected mean values could be initialized either by preselected values or in random. Then, all points are set to the cluster with minimal distance to its centroid. After the points are distributed to clusters, the values of $m^{(k)}$ are recomputed and updated. Then in case of need, all points with adjusted minimal distances to the new centroid values are replaced to the newly determined clusters. The process proceeds until no assignments are available. It is a finite process because the algorithm always converges to a fixed point (MacKay, 2003).

The outcome of the algorithm depends on the initial conditions and the characteristics of the initial data set are very important. The algorithm is not very helpful when expected groups have a specific shape, position or their populations are very different in size. In these cases, the outcome is very difficult to be estimated and usually is wrong. Other difficulties in implementing K-means clustering may arise due to data distribution, especially in case of the existence of overlapping area between clusters. In this case, when the number of shared points is small, their careful removal could be useful. But, in case of the removal of a large number of similar points, the model overfit could be achieved very easily.

Another very important input parameter is the initial number *k* of expected clusters. This number could be preselected by prior knowledge, similarly to some previous works (Piloyan and Konečný, 2017). However, this approach is not applicable without a profound knowledge of the distribution of observed data. Likewise, such presumption might not yield an optimal clustering of the available data, resulting in a lack of global optimum. Finally, it is selected a numerical procedure to determine automatically the optimal number of clusters *k*. For this purpose, the algorithm must be run for a range of $1{:}K_{max}$ possible number of clusters and select the optimal value based on numeric criteria.

The most popular criteria to determine *k* is the total within-cluster sum of square (wss). The optimal number of centroids is located on the bend (knee) of the curve of values of wss in range $1{:}K_{max}$. Because the overall variations always decrease with expanding the number of clusters, the decision is based on visual detection the value of *k* for which the rate of decrease sharply shifts. This makes the method ambiguous in many cases. For this reason, the average silhouette method is selected as an alternative criterion (Rousseeuw, 1987). It is a ratio scale of how similar an object is to its own cluster compared to other clusters. The silhouette ranges from −1 to +1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighbouring ones. Thus, the optimal number of clusters *k* is determined by the highest value of average ratios of all points (average silhouette).

Finally, the clustering algorithm is supposed to be implemented simultaneously on multivariate data. But the parameters have different scales with very big differences. This imposes domination of the values with higher mean values over these with lower, due to strong dependency on initial conditions. To avoid similar variable variations, the input data is standardized prior to clustering (Davis, 2002). Moreover, the transformation reduces the impact of outliers and allows comparing a sole observation against the mean.

There are many implementations of K-means clustering for almost all mainly used programming languages and scripts. However, it is developed a special script for R ({R Core Team}, 2019). It executes consecutive procedures for automated computation of different cluster options. The program is an envelope over the available implementations of K-means cluster in R. The mainly used function is *kmeans* from the *stat* packets of R. For determination of the optimal number of clusters are used functionalities of the Factoextra R Package (Kassambara and Mundt, 2017). Finally, for computations of geospatial functionalities are libraries related to *sp* package (Bivand et al., 2013).

3. RESULTS

The computations of clustering with morphological data of the Ogosta Valley are automatized and quick. However, the estimation for validity and agreement to the reality is not so straightforward because of the lack of predefined classification, despite some intuitive assumptions for the existence of a distinct cluster upstream the village of Beli Mel and the separation of theactive river floodplain from this part of the valley bottom which is more distant from the river. For this reason, the verification procedure relies mainly on the resilience of acquired clusters to changes and their compliance with the geomorphographic units of terrain, and the extent and depth of the flood in 1964.

Multiple numbers of cases with a different combination of input data and parameters are observed. The parameters used together or in a different combination, are the *distance to the river*, *VDCN*, *MRVBF*, *DTM* and *mTWI*. The algorithm in the initial step is run for all parameters over the complete dataset. Then, for further development, additional computations with different combinations of parameters and subsets of data points are performed. The selection of data and parameter combinations is determined by previously acquired results. In some cases, a minor number of data points which are not clearly assigned to any centroid are additionally removed from the data set. The outcome is improved clustering and clarified cluster properties. Because of the large number of computed case scenarios, they are given classification names which are shown in *italics letters* further in the text.

3.1. Initial clustering

The clustering process commenced for the complete data set and with all five parameters. This model is referenced as *initial clustering*. The optimal number of clusters, in this case, is yielded to four by the silhouette method (Table 3). The model clearly divides the valley into two parts with different geomorphological characteristics, which is one of its main advantages. The narrow stretch of the valley upstream the settlement of Beli Mel is occupied by the *cluster 1* (Figure 2).

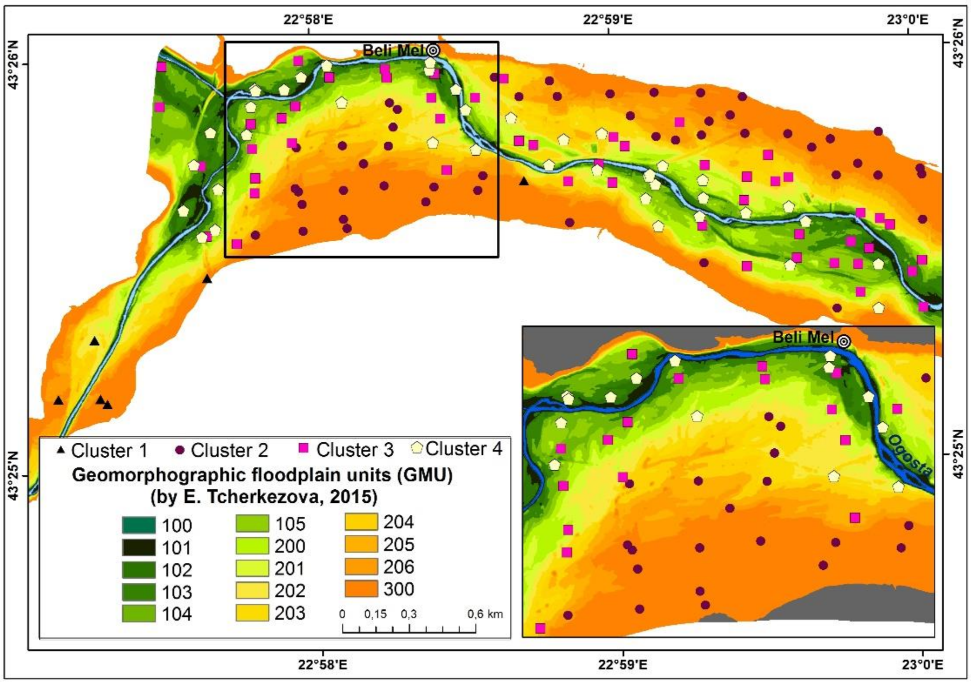

Figure 2. Spatial distribution of the clusters computed with all predefined parameters and data

The other three clusters are mainly located downstream of the village, where the valley floor becomes much wider. *Cluster 2* consists of sites located further from the river and higher above its channel. The data points of cluster 4 tend to be close to the river, while *cluster 3* is allocated a little further from the river banks. However, the two groups of data points are less clearly separated within the valley, compared to *clusters 1* and *2*. However, four data points are distinguished and may be classified as spatial outliers. They are estimated as notably different by two different classifications performed in the next steps of the analysis. Three of the points are located in the area upstream the settlement Beli Mel. They are classified as belonging to *clusters 3* and *4*, due to their location very close to the river. The distinction between the three points is mainly a result of very high *MRVBF* and differences in *VDCN*. The last spatial outlier is a point belonging to *cluster 1* but located in the area downstream the Beli Mel. The main reason for inclusion in *cluster 1* is the combination of a very high value of the *VDCN* and the short distance to the river. These cases show the sensitivity of the model.

To verify that the obtained differences between computed clusters are mainly related to the geomorphographic characteristics of the valley, a comparative analysis on parametric distribution is applied in addition to spatial analysis. The distribution of the data points of each cluster between the *GMUs* display clear differences between the obtained clusters, as it is shown in Table 4 and Figure 3.

Table 3. The mean values and standard errors (in brackets) of the geomorphological parameters calculated by clusters.

	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**
MRVBF	1.30 (± 1.00)	2.04 (± 1.26)	**3.70 (± 0.89)**	1.04 (± 0.82)
VDCN	**5.20 (± 2.25)**	**6.63 (± 1.90)**	3.00 (± 1.12)	2.59 (± 1.20)
DTM	**392.1 (± 63.9)**	241.0 (± 30.5)	226.5 (± 31.1)	229.8 (± 33.6)
mTWI	3.70 (± 1.20)	4.66 (± 1.12)	5.15 (± 0.66)	3.55 (± 1.1)
Distance to Ogosta	54.87 (± 38.9)	**338.8 (± 143)**	165.9 (± 102)	85.36 (± 83.9)

Cluster 1 is distinctly separated in the mountainous part of the valley upstream the Beli Mel. This cluster is the one with the highest altitudes. *Cluster 2* populates mostly the upper floodplain (T_{0-h}) and the periphery of the valley floor downstream the village of Beli Mel. This area is dominated by the GMUs 200-300 with an elevation higher than 3 m above the river. *Clusters 3* and *4* occupy primarily the active floodplain (T_{0-l}) in the same part of the valley where GMUs 102-105 are well developed at an elevation lower than 3 m. The difference between the two subsets can be seen in the predominance of the sites located in the bankfull channel zone (GMUs 100-101) for the *cluster 4*, and in the upper floodplain for the *cluster 3* (Table 4).

The obvious conclusions of differences can be easily confirmed with the Chi-square test for independence between the cluster groups. The result shows extremely low probabilities of any similarities on *GMU* distribution between clusters (the Chi-square is equal to 332.16 with 9 degrees of freedom).

Table 4. Distribution of the data points of each cluster between the geomorphographic units

Morphographic units	**Cluster 1** **Numbers**	**Cluster 2** **Numbers**	**Cluster 3** **Numbers**	**Cluster 4** **Numbers**
100-101	4	0	10	23
102-105	19	3	175	116
200-206	40	84	80	33
300	19	58	0	1
Total	**82**	**145**	**265**	**173**

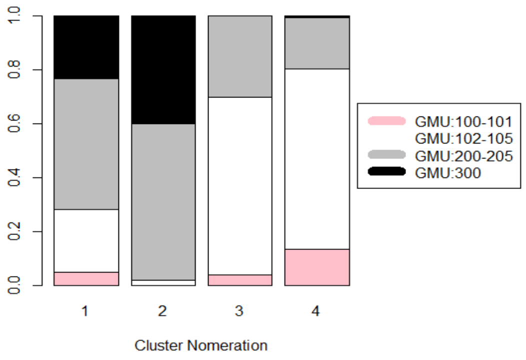

Figure 3. Bar chart of GMU frequency distribution over all 4 clusters

Additional computations with selected parameters are performed to estimate their particular impact on the clustering model. When the parameters *VDCN* and *MRVBF* are only used, the data set is clearly split by the position above the river channel. In this case, the optimal solution is a clustering of 3 groups of data points which are centred over the river with a difference in the height above the river bed. However, all scenarios yield clusters without clear distribution by *distance to the river*. The inclusion of *mTWI* only interchanges a small number of points between clusters and neither improves the spatial distribution nor clears the separation by parameters in clustering. When the *distance to the river* is added to *VDCN* and *MRVBF*, the number of clusters is shrunk to 2, but the difference between them shifts mainly to the *distance to the river*.

The cluster model with *DTM*, *VDCN* and *mTWI* distinguish 4 clusters concerning terrain irregularities. The *DTM* turns to be a very important contributor to the spatial distribution of the data points, projecting the serious decrease of the absolute terrain altitudes of the valley floor downstream the river course. Thus, with the inclusion of the *DTM,* the model begins to follow not only the relative position to the river determined by the distance to and the height above the river channel but also the altitude differences across the whole valley. The addition of the *distance to the river* to the model orientates it to the river floodplain. Finally, the inclusion of the *mTWI*only improves the precision of the inter-cluster borders, but do not change the main differences between clusters.

3.2. Second clustering

In order to refine the *clusters 2, 3* and *4*, the data points upstream the village of Beli Mel are removed and the analysis is performed only with the sites located in the wider section of the valley, excluding the spatial outliers of *clusters 3* and *4*. Regarding the spatial outlier of *cluster 1,* it does not impact the following computations and it is excluded for simplicity from overall statistics. Thus, the dataset is reduced to 581 different sites and the performed clustering with this data will be referenced as *second clustering* in the following text. The new run of K-means clustering with all five parameters confirmed the optimal number of clusters equal to 3. A small number of points shifted from and to *cluster 2*. The leading parameter by which points are mostly changed in-between clusters is the vertical distance. The sites with a lower altitude are replaced by the ones with higher elevation from *cluster 3* and *cluster 4.* For the rest parameters, the mean values remained with minor changes. The resulting new *GMU* distribution of adjusted groups is shown in Table 5.

Table 5. GMU distribution for K-means clustering with 3 clusters. The changes related to every group from the previous distribution are shown in brackets. The changes are revisited for removed samples.

GMU Range	**Cluster 2** **Numbers**	**Cluster 3** **Numbers**	**Cluster 4** **Numbers**
100-101	0 (-)	10 (-)	22 (-)
102-105	0 (-3)	171 (-3)	122 (+6)
200-206	88 (+4)	72 (-8)	37 (+4)
300	58 (-)	0 (-)	1 (-)
Total	**146(+1)**	**253 (-11)**	**182 (+10)**

Despite the changes, the profiles of obtained clusters from *initial clustering* are preserved. Even more, the changes clarified distributions sending some of the points to more appropriate groups. For instance, all the three sites available in *cluster 2* and located in the active floodplain (*GMU* 102-105) are substituted with other ones from the upper floodplain (*GMU* 200-206). This distribution is confirmed with the results from the simulated flood event in 1964. Because *cluster 2* mainly consists of sites which are remote and high above the river, the expected risk of flooding is very low. Only one site from this group is found within the flood area with an inundation depth of almost zero. Conversely, the *clusters 3* and *4* are intensively flooded across the valley (Figure 4).

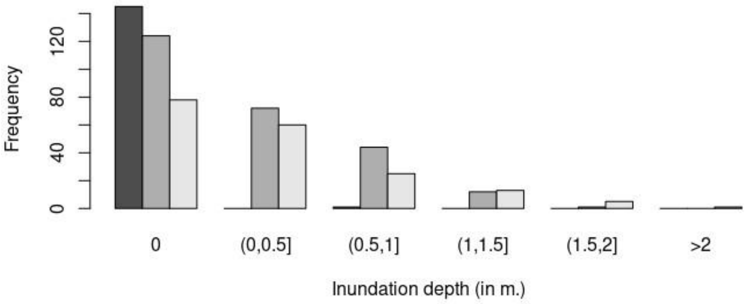

Figure 4. Combined histograms of the estimated inundation depth during the flood in 1964 for the cluster 2 (black), cluster 3 (dark grey) and cluster 4 (light grey).

3.3. Third clustering

There are several reasons for additional classification for *cluster 3* and *cluster 4*. Firstly, they cannot be spatially separated very clearly because the two groups of data points share a common area in the floodplain. Secondly, significant shares of the sites of both clusters are inundated in 1964 according to the flood simulation, which is evidence that they populate similar areas in the active floodplain.

The new clustering is run for two centroids which are predefined from the previous two models. It is performed on the sub-dataset consisted of the points of *clusters 3* and *4* computed in the *second clustering*. The outliers determined for the first model are not considered in the analysis. The parameters *MRVBF*, *DTM* and *mTWI* are used in the third scenario.

The new clustering almost repeats the initial results. Only nine sites changed their classification in-between both clusters. They interchanged their groups only to improve precision, but without to cause profile changes. As a result, the points are clearly grouped by *MRVBF*, which reflects the terrain irregularities (Figure 5).

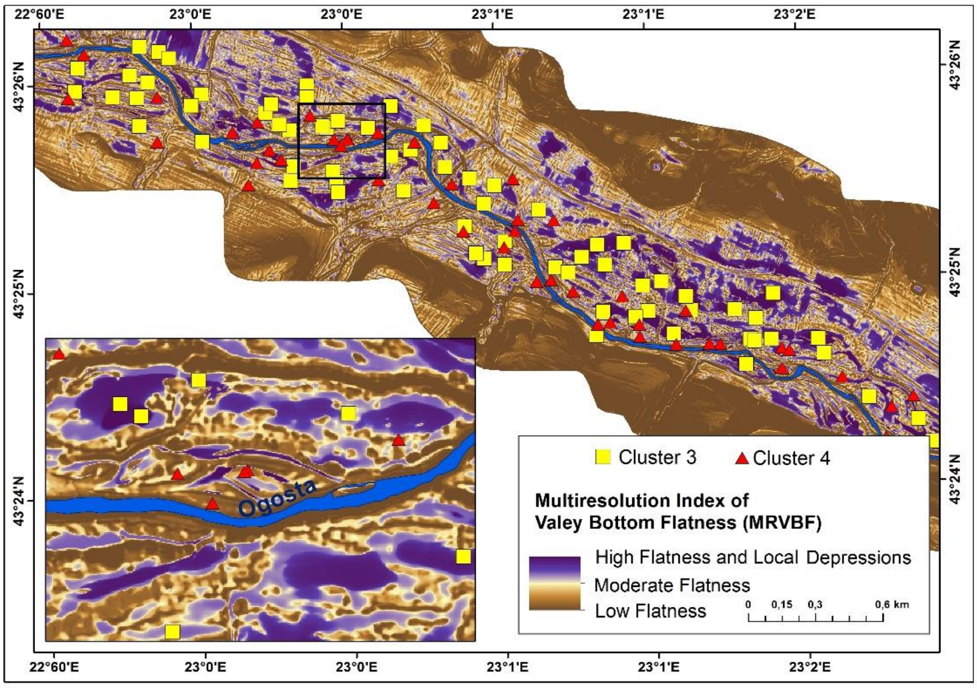

Figure 5. Distribution of clusters 3 and 4 obtained by scenario for two centroids, and MRVBF, DTM and mTWI, presented on the map of valley bottom flatness

However, the silhouette method shows that the partition of the selected data sample on three parts is more optimal. A new model is elaborated for three centroids and for the same parameters *MRVBF*, *DTM* and *mTWI*. We name this model as *third clustering*. Likewise, despite the minimal gained optimizing effect, the reshuffling of points to three clusters make a big difference. Most of the sites of *cluster 3* (200 of 253) with high *MRVBF* values remained in the newly computed *cluster 5-1*. Almost all remaining sites in *cluster 3* (50 from 53 remained) are included in the intermediate *cluster 5-2*. Conversely, the sites of *cluster 4* populate mainly *cluster 5-2* (72 of 171) and *cluster 5-3* (98 of 171). This new cluster configuration, similarly to the previous variant, splits points mainly by the valley bottom flatness (Figure 6). However, the new configuration does not produce clear spatial differentiation between *clusters 5-2* and *5-3* within the floodplain, as well as regarding the depth of the simulated inundation.

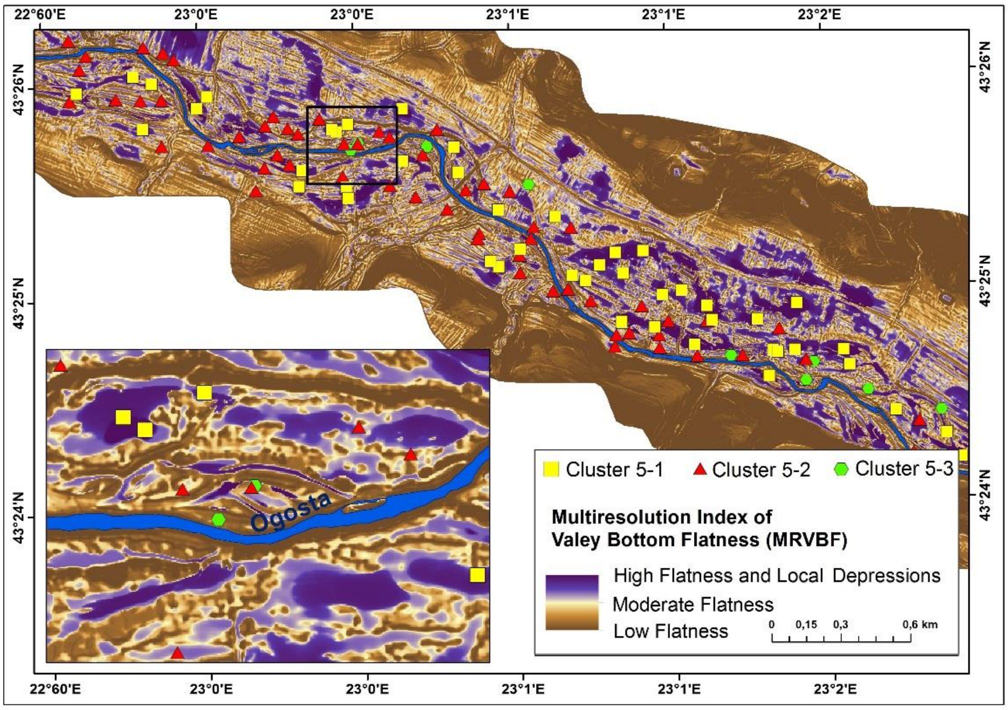

Figure 6. Distribution of clusters 5-1, 5-2 and 5-3 on the map of the valley bottom flatness

3.4. Stability and overfitting of the clustering models

All obtained models are tested for overfitting and stability. For this purpose, every clustering model is recomputed more than 1000 times with different sample splitting. The samplings are produced with a random selection of a ¾ part of data points in every cluster after independent trials. The criteria for stability are the reproducibility of the recomputed optimal number of centroids and the cluster classification of every data point after every repetition. The results of the tests confirm the stability of the obtained model with 4 clusters yielded from the *initial clustering* **-** the optimal number of clusters of four is yielded in about 85% of tests and the number of interchanging points is below 5% for every cluster.

However, the repeated computations of *the third clustering* show very strong variability – the optimal number of clusters is confirmed in less than 70% of tests and data variability is above 40% for every cluster. The main reason for this instability is the decrease in the significance of *DTM* and *VDCN* on the reduced data set. Thus, the model degrees of freedom are reduced and the newly computed models have more parameters that can be justified by data. This classifies the process of multiple repetitions of K-Means clustering as not suitable for improvement by reduction of parameters.

4. CONCLUSIONS

This paper presents an approach for grouping the data points of a soil sampling network to support the spatial analysis of metal contamination of soil in a fluvial environment. K-means clustering is used for grouping the available points by five morphographic and morphometric variables, e.g. altitude, distance to the river, vertical distance to channel network, multiresolution index of valley bottom flatness, and modified topographic SAGA wetness index. It results in grouping of the sampling sites in the Ogosta Valley into four parts which are attached to distinct areas in the valley floor with specific sedimentary environment. The altitude has the most significant impact, dividing the valley floor into two morphologically contrasting sections: a narrow mountainous part with well-developed upper floodplain, and a wide section downstream of it, with a broad lower (active) floodplain. The soil sampling sites in the wider part of the valley are well distinguished between the upper and lower floodplains according to their distance to the river channel and the vertical distance to it. The two clusters located in the active floodplain differ mostly by the values of the valley bottom flatness but are not so clearly separated in the space. The attempts to do some new grouping in-between the two clusters make the clustering model unstable.

The applied algorithm for grouping and classification of big-sized geomorphological data with K-means clustering is fast and easy to compute, as it requires up to four computational iterations. The outcome is found to be in good agreement to the general properties of the terrain. The model is consistent with the previously determined geomorphographic units in the valley, and with the extent and inundation depth of a simulated historic flood event. However, the effectiveness is strongly dependent on the appropriate selection of the parameters and their independent and regular spatial distribution. In case of the existence of dependency in data, correction must be applied to the selected distance measure. Repeated usage of K-means clustering on already classified groups may be useful for better understanding and discovery of hidden patterns and properties. However, the results must be taken very carefully because of the risk of overfitting.

The present algorithm can be applied for producing more homogeneous statistical samples of soil sampling sites to find regression models of the spatial distribution of the heavy metal concentrations in soil, depending on the topography of the river valleys.

ACKNOWLEDGEMENTS

This study was supported by the National Science Fund of Bulgaria, Grant No ДН 04/3, project ARSENT. The survey also benefited from the project ASCOR which was funded by the Bulgarian-Swiss Research Programme 2011-2016.

The authors of the article would like to sincerely thank Ruben Kretzschmar and Kurt Barmettler from the Institute of Biogeochemistry and Pollutant Dynamics, ETH Zürich, Switzerland for their help and assistance in conducting this study.

REFERENCES

Arnold, J.G., Kiniry, J.R., Srinivasan, R., Williams, J.R., Haney, E.B, and S.L. Neitsch (2012). *Soil & Water Assessment Tool. Input/Output Documentation, Version 2012.* Texas Water Resources Institute. TR-439, 654 p. https://swat.tamu.edu/media/69296/swat-io-documentation-2012.pdf (Accessed 2019-02-07).

Beechie, T., and H. Imaki (2014). Predicting natural channel patterns based on landscape and geomorphic controls in the Columbia River basin, USA, *Water Resource Res* (50): 39–57, doi:10.1002/2013WR013629

Bird, G., Brewer, P., Macklin, M., Nikolova, M., Kotsev, T. and Swain, C. (2010). Dispersal of Contaminant Metals in the Mining-Affected Danube and Maritsa Drainage Basins. Bulgaria, Eastern Europe. *Water Air Soil Pollution.* 206 (1): 105-127.

Bivand, R.S., Pebesma, E., and Gómez-Rubio, V. (2013). *Applied spatial data analysis with R*, Second edition. Springer, NY., XVIII, 405 p., ISBN 978-1-4614-7618-4.

Bock, M., Boehner, J., Conrad, O., Koethe, R., and A. Ringeler, A. (2007). *Methods for creating Functional Soil Databases and applying Digital Soil Mapping with SAGA GIS*. In: Hengl, T., Panagos, P., Jones, A., and G., Toth (eds.). *Status and prospect of soil information in south-eastern Europe: soil databases, projects and applications*. EUR 22646 EN Scientific and Technical Research series, Office for Official Publications of the European Communities, Luxemburg, p. 149-162. http://eusoils.jrc.ec.europa.eu/ESDB_Archive/eusoils_docs/esb_rr/EUR22646EN.pdf (Accessed 2019-02-07).

Brewer, P., Dennis, I. and Macklin, M. (2005). The use of geomorphological mapping and modelling for identifying land affected by metal contamination on river floodplains. Defra Research and Development Report SP 0525.

Brierley, G. and K. Fryirs (2000). River Styles, a Geomorphic Approach to Catchment Characterization: Implications for River Rehabilitation in Bega Catchment, New South Wales, Australia. *Environmental Management* (25): 661, https://doi.org/10.1007/s002670010052.

Bruner G.W. (2010). *HEC-RAS River Analysis System. Hydraulic Reference Manual, Version 4.1. January 2010. Approved for Public Release*, CPD-69, 411 p. https://www.hec.usace.army.mil/software/hec-ras/documentation/HEC-RAS_4.1_Reference_Manual.pdf (Accessed 2019-02-07).

Ciszewski, D., Urszula Kubsik, U. and Aleksander-Kwaterczak, U. (2012). Long-term dispersal of heavy metals in a catchment affected by historic lead and zinc mining. *Journal of soils and sediments* 12: 1445-1462.

Clement, A., Novakova, T., Edwards, K., Fuller, I., Macklin, M., Fox, E. and Zapico, I. (2017). The environmental and geomorphological impacts of historical gold mining in the Ohinemuri and Waihou river catchments, Coromandel, New Zealand. *Geomorphology* 295: 159-175.

Conrad, O., Bechtel, B., Bock, M., Dietrich, H., Fischer, E., Gerlitz, L., Wehberg, J., Wichmann, V., and J. Boehner (2015). System for Automated Geoscientific Analyses (SAGA) v. 2.1.4. G*eosci. Model Dev.*, 8, 1991-2007, doi:10.5194/gmd-8-1991-2015

Davis, J. (2002). *Statistics and data analysis in geology*-3'd ed. John Wiley & Sons, Inc

Dennis, I., Macklin, M., Coulthard, T. and Brewer, P. (2003) The impact of the October–November 2000 floods on contaminant metal dispersal in the Swale catchment, North Yorkshire, UK. *Hydrological Processes* 17: 1641–1657.

Dunning, S., Massey, C., and N. Rosser (2009). Structural and geomorphological features of landslides in the Bhutan Himalaya derived from terrestrial laser scanning. *Geomorphology* (103): 17–29.

Gallant, J., and Dowling, T. (2003). A multiresolution index of valley bottom flatness for mapping depositional areas. *Water Resources Research* 39 (12): 1347, doi:10.1029/2002WR001426.

Jordanova, D., Goddu, S.R., Kotsev, T., Jordanova, N. (2013). Industrial contamination of alluvial soils near Fe-Pb mining site revealed by magnetic and geochemical studies. *Geoderma* 192: 237-248.

Kassambara, A. and Mundt, F. (2017). *factoextra: Extract and Visualize the Results of Multivariate Data Analyses*. (Accessed 2019-01-15)

MacKay, D. (2003). *Information Theory, Inference and Learning Algorithms*. Cambridge University Press.

Macklin, M.G. Brewer, P.A., Hudson-Edwards, K.A., Bird, G., Coulthard, T.J., Dennis, I.A., Lechler, P.J., Miller, J.R., and J.N. Turner (2006). A geomorphological approach to the management of rivers contaminated by metal mining. *Geomorphology* 79 (3-4): 423–447, doi: 10.1002/aqc.467.

Mandaliev, P., Mikutta, Ch., Barmettler, K., Kotsev, Ts., and R. Kretzschmar (2014). Arsenic species formed from arsenopyrite weathering along a contamination gradient in circumneutral river floodplain soils. Environmental Science & Technology 48(1): 208–217, doi: 10.1021/es403210y.

Martin, J.-M., and M. Meybeck (1979). Elemental mass-balance of material carried by major world rivers. *Marine Chemistry* 7 (3): 173–206.

Piloyan, A. and Konečný, M. (2017). Semi-Automated Classification of Landform Elements in Armenia Based on SRTM DEM using K-Means Unsupervised Classification. *Quaestiones Geographicae* 36 (1): 93-103, doi: 10.1515/quageo-2017-0007.

R Core Team, R Foundation for Statistical Computing (2018). *R: A Language and Environment for Statistical Computing*. https://www.R-project.org/ (Accessed 2018-12-28).

Rousseeuw, P.J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics,* 20: 53-6.

Simmler, M., Suess, E., Christl, I., Kotsev, T., Kretzschmar, R. (2016) Soil-to-plant transfer of arsenic and phosphorus along a contamination gradient in the mining-impacted Ogosta River floodplain. Science of Total Environment 572:742-754, ISSN:0048-9697.

Spectroteh. (1994). Establishment of the Type and Degree of Ecologically Polluted Farmland with Heavy Metals in the Municipalities of Chiprovtsi and Georgi Damyanovo (in Bulgarian): Sofia, Spectroteh report, p. 33.

Taylor, M. and Hudson-Edwards, K. (2007). The dispersal and storage of sediment-associated metals in an arid river system: the Leichhardt River, Mt Isa, Queensland**.** *Environment Pollution* 152 (1): 193-204.

Tcherkezova, E. (2015). GIS-based delineation and regionalization of geomorphographic units in the floodplain of Ogosta River between the settlement Gavril Genovo and the “Ogosta” reservoir (NW-Bulgaria). *Problems of Geography,* 1-2, Academic Publishing House “M. Drinov: 114-122.

Thomson, J.R., Taylor, M.P., Fryirs, K.A., and G.J. Brierley (2001). A geomorphological framework for river characterization and habitat assessment. Aquatic Conservation: Marine and Freshwater Ecosystems 11(5):373-389, https://doi.org/10.1002/aqc.467

Valentine, A. and L. Kalnins (2016). An introduction to learning algorithms and potential applications in geomorphometry and Earth surface dynamics. *Earth Surf. Dynam.* (4): 445-460, https://doi.org/10.5194/esurf-4-445-2016.

Weis, K. and Hronček, P. (2017). Using historic postcards and photographs for the research of historic landscape in geography and the possibilities of their digital processing. European Journal of Geography 8 (5):77 –85.