Creating Maps Defining the Level of Acidification by Fuzzy Analysis

Michal Gréc\textsuperscript{1,*}, Matěj Horáček\textsuperscript{2}, Vilém Novák\textsuperscript{3}

\textsuperscript{1}Department of Mathematics, Faculty of Science, University of Ostrava, Czech Republic
\textsuperscript{2}Department of Physical Geography and Geocology, Faculty of Science, University of Ostrava, Czech Republic
\textsuperscript{3}Centre of Excellence IT4Innovations, Division of the University of Ostrava, Institute for Research and Applications of Fuzzy Modeling, Czech Republic

Email: michal.grec@osu.cz, matej.horacek@osu.cz, vilem.novak@osu.cz

\textsuperscript{*}Corresponding author.

\textbf{Abstract}
Acidification is one of the most important chemical properties of soil and an important indicator of soil quality. Long-term acidification and cation nutrient degradation of forest soils is proving to be the major factor that will limit forest management in the near future. We analyze soil horizons. We use classical cluster analysis and fuzzy cluster analysis to identify areas that have been affected by acidification. The goal of the work is to determine the degree of damage and creating RGB map that would show areas with the same or similar damage caused by acidification. Part of the work is also searching for the areas which have totally different level of acidification (critical points). The result is a map that was created by mixing RGB colors.

\textbf{Keywords}
Acidification — Fuzzy analysis — Clustering analysis — RGB map — Soil damage.

\textcopyright 2015 by Orb Academic Publisher. All rights reserved.

1. Introduction
Accumulated dirt of forest soil sand and their degradation due to acid deposition just begun to fully show in the Czech Republic [1]. Soil acidity is among the characteristics that are often considered in the evaluation of soil quality [1]. Soil horizon is a layer of soil that has a specific horizontal location of a certain physical and chemical properties. It is a layer of soil that has identical external characteristics (color and structure) and the same physical characteristics and chemical composition.

Damage to forest stands and their decline in the Czech Republic evoked a necessity to elucidate the function of soil in these processes. Main factors affecting acidification of forest soils are assessed in the paper.

The soil plays an important role within forest ecosystems because it affects resistance and resilience to a considerable extent and thus the stability of the whole system [2].

In real applications, boundary between clusters is not sharp very often, thus fuzzy clustering is better suited for the data. The goal of this work is to create RGB map of local forest area in the Czech Republic.

Maps should represent areas of soil horizons influenced by acidification and degree of damage. In the second section we describe process of acidification, (causes and consequences) and section III describes three variables used in our calculations. In section IV, we discuss cluster analysis and fuzzy algorithm Fanny. We chose fuzzy cluster analysis for more variable description of acidification. In section V we discuss about packages used in R, which is a language and environment for statistical computing and graphics. In the last section obtains description of creating RGB maps. Soil acidity is a major land degradation issue facing much of Czech Republic. It is important to know if it is being managed appropriately to avoid costly losses in production or negative implications for the environment.

2. Acidification
Acidification of soils is generally a result in the formation of acid soils or their feed from the outside. Acidification of the soil is defined also as a decrease in the neutralization capacity of the soil [1].

Soil acidification is a process by which soil pH decreases over time, and there are often no visible signs of the problem. Acidification can occur under natural conditions over thousands of years, with high rainfall areas most affected. However, rapid acidification can occur over a few years under intensive agricultural practices.
Acidification can affect either the surface soil only or the subsoil as well. Surface acidity can be relatively simple to treat, and brings considerable benefits in plant growth and yield. Subsurface acidity is difficult and costly to correct. And that is a reason why we need to identify the problem as early as possible especially in high-risk areas [3].

This process is a result of formation of organic acids which occur in forest soil by decomposition of organic substances, especially litter and surface humus.

Emerging organic acids bind to each alkaline cations - ions of calcium, magnesium and potassium, which are highly mobile in soils and which are actively involved in the process of neutralization of organic acids. The resulting compounds - organic salts - are in conditions of excessive precipitation washed out of the soil, thereby causing its acidification [1].

Acidification of forest soils may have the following sources [4]:

- Supply of dissolved strong acids and alkalies.
- Internal production of various acids in soil.
- Assimilation of basic substances by biota.
- Changes during the reduction-oxidation processes.

There are no visible symptoms of soil acidification. As soils become more acidic some nutrients may become less available while other elements in the soil may reach toxic levels. Acidic soils may have some or all of the following problems:

- Reduction in the amount of nutrients being recycled by soil micro-organisms (e.g. nitrogen supply may be reduced).
- The ability of plants to use subsoil moisture may be limited [3].

3. Measured Variables

The soil itself can be regarded as a relatively stable component of forest ecosystems considering the long-term character of its genesis [2].

We work with the data that form the matrix of size 817x6 of certain local area. Our data are selected pedochemical characteristics of associated soil horizons in the network 1x1 km created for the forest lands in the Czech Republic. We have three attributes S (saturation of alkalies), C (concentration of basic cations) and E (content of exchangeable alkalies) in two horizons uh (upper horizon - organo-mineral horizon mostly black, lying just under the surface) and dh (diagnostic horizon - lies below the upper horizon is the most powerful (thickest), usually brown color).

The rate and intensity of soil processes have already been the subject of interest of soil scientists for a long time because the rate and intensity not only explain potential anthropic impacts on these processes in the past but also they are a condition for elaborating the prediction of potential consequences of human activities in the development of soil properties closely related to forest soil fertility [2].

We made the transformation for determining the same range of data. The values are expressed in millimoles of the chemical equivalent (mmol) on 0.1kg (100g) of soil. The values of this indicator change during the year, changes in soil moisture and fertilization [5].

Saturation of alkalies (S) indicates the proportion of neutral calcium cations of magnesium, potassium and sodium on the total effective capacity of the cation exchange, while the rest of acidic cations (H, Fe, Al, Mn, Ma) is received [6].

Saturation of alkalies is the decisive indicator that describes the acidity of soil. Data of the saturation of alkalies do not need to be transformed. Their initial values are already in the range 0-100 (in %).

Cation exchange capacity (C) characterizes the overall ability of the soil to bind cations (cmol+/kg).

\[ x = \log_{1,2}(x_i + 1) \times 4 \] (1)

Content of exchangeable alkalies (E) is the number of alkalies that is bound just the sorption complex. (cmol+/kg)

\[ x = (\frac{x_i}{2}) \times (100 ÷ 3,33) \] (2)

4. Cluster Analysis

The primary goal of cluster analysis is to classify objects into groups (clusters), mainly so that the two objects of the same cluster are more similar than two objects from different clusters. Cluster analysis divides data into groups (clusters) that are meaningful, useful, or both [7].

Cluster analysis groups data objects based only on information found in the data that describes the objects and their relationships. The goal is that the objects within a group be similar (or related) to one another and different from (or unrelated to) the objects in other groups.

The greater the similarity (or homogeneity) within a group and the greater the difference between groups, the better or more distinct the clustering [7]. Each object is uniquely assigned to one cluster. The basic division of clustering methods according to the destination to which they are headed is the hierarchical and non-hierarchical [8].

Hierarchical clustering is a system of subsets, where the intersection of two subsets - clusters is the empty set or one of them.

Non-hierarchical clustering is a system, where the intersection of clusters is empty. They are disjoint sets. The basic concept of clustering analysis is the distance between objects.

It can be the absolute value of the difference, Euclidean distance, or any other function \( d : V \times V \rightarrow R \), which has the following characteristics:

\[ d(x,x) = 0, \]
\[ d(x,y) = d(y,x), \]
\[ d(x,z) \leq d(x,y) + d(y,z) \]

for all \( x,y,z \in V \).

The most often used metrics are:
• **Euclidean metric** it is probably the most used metric. The calculation is based on the Pythagorean theorem.

\[
d_E(x_k, x_l) = \sqrt{\sum_{j=1}^{p} (x_{kj} - x_{lj})^2}
\]  

(3)

• **Manhattan metric** (city-block metric) in many cases leads to similar results as the Euclidean distance.

\[
d_{CB}(x_k, x_l) = \sum_{j=1}^{p} |x_{kj} - x_{lj}|
\]  

(4)

• **Minkowski metric**

\[
d_M(x_k, x_l) = \sqrt{\sum_{j=1}^{p} (x_{kj} - x_{lj})^z}
\]  

(5)

for \( z = 1 \) is a Manhattan metric and for \( z = 2 \) is the Euclidean metric. The higher is \( z \), the greater is difference between objects.

• **Mahalanobis metric.** This metric includes bonds between characters expressed covariance matrix \( C \). It is about the distance of points in space whose axes don’t need to be orthogonal.

\[
d_Ma(x_k, x_l) = \sqrt{(x_k - x_l)^T C^{-1} (x_k - x_l)}
\]  

(6)

The problem of all distance metrics arises when using non-standardized data, which may cause differences between the clusters (difference measurement units) [9].

From the mathematical point of view, cluster analysis is searching for suitable decomposition of \( V \), i.e. finding such a system of subsets (clusters) \( \{R_j|j = 1, \ldots, c\} \) et \( R_j \subset V \) for which the following applies:

\[
\bigcup_{j=1}^{c} R_j = V,
\]

\[
R_i \cap R_j = \emptyset, i \neq j, 1 \leq i, j \leq c
\]  

(7)

and which also satisfies the requirement that the elements in one cluster \( R_j, j = 1; \ldots; c \) were each "closer" than elements from different clusters. In fuzzy analysis clustering are \( R_j \subset V \) fuzzy sets [10]. Also, the fuzzy sets must fulfill the following conditions: let \( V = \{v_1; \ldots; v_r\} \); then

\[
\sum_{j=1}^{c} R_j(v_i) = 1, \quad i = 1, \ldots, r
\]  

(8)

\[
0 < \sum_{j=1}^{c} R_j(v_i) < r, \quad j = 1, \ldots, c
\]  

(9)

The first condition says that every element of \( V \) must belong to at least one cluster and the sum of all degrees of membership in all clusters must be equal to 1.

The second condition says that all the elements of \( V \) must not belong in the maximum degree membership to any cluster and each cluster must be non-empty [11].

In classical cluster analysis each datum must be assigned to exactly one cluster. Fuzzy cluster analysis relaxes this requirement by allowing gradual memberships, thus offering the opportunity to deal with data that belong to more than one cluster at the same time.

Cluster analysis deals with the discovery of structures or groupings within data. Since hardly ever any disturbance or noise can be completely eliminated, some inherent data uncertainty cannot be avoided. That is why fuzzy cluster analysis dispenses with unambiguous mapping of the data to classes and clusters, and instead computes degrees of membership that specify to what extent data belong to cluster [11].

Most fuzzy clustering algorithms are objective function based: They determine an optimal classification by minimizing an objective function.

The degrees of membership, to which a given data point belongs to the different clusters, are computed from the distances of the data point to the cluster centers with regard to the size and shape of the cluster as stated by the additional prototype information. The closer a data point lies to the center of a cluster (size and shape), the higher is its degree of membership to this cluster [12]-[13].

In fuzzy clustering, each observation can belong to more clusters. Denote by \( u(i, v) \) the membership of observation \( i \) to cluster \( v \). We decided to use algorithm \( \text{fanny} \), which computes a fuzzy clustering of the data into \( k \) clusters:

\[
fanny(x, k, \text{diss = inherits(x, "dist"), memb.exp = 2, metric = \text{c(“euclidean”, “manhattan”, “SqEuclidean”)), stand = FALSE, iniMem.p = NULL, cluster.only = FALSE keep.diss = !diss && !cluster.only & & n < 100, keep.data = !diss & & !cluster.only, maxit = 500, tol = 1e-15, trace.lev = 0})
\]

where \( x \) is a value, \( k \) is a number of clusters (in our case it is 3), \( \text{diss} \) is logical flag (T/F) and \( \text{memb.exp} \) is membership exponent.

\( \text{Fanny} \) aims at minimization of the objective function

\[
\sum_{v=1}^{k} \frac{\sum_{j=1}^{n} \sum_{i=1}^{r} u_{iv}^p d(i, j)}{2 \sum_{j=1}^{k} u_{iv}^p}
\]  

(10)

where \( n \) is the number of observations (in our case tenth of the data), \( k \) is the number of clusters (3 – R, G, B), \( r \) is the membership exponent, that influences the fuzziness of the clustering (the default membership exponent is set to 2, but this has been reduced on 1.5 to obtain a less fuzzy membership of objects in each cluster) and \( d(i, j) \) is the dissimilarity between observations \( i \) and \( j \) [14].

\[
d(i, j) = \sqrt{\sum_{k=1}^{n} (i_k - j_k)^2}
\]  

(11)
Note that $r \to 1$ (if $r$ approaches 1 then we obtain more crisp clusters while if $r$ approaches infinity then we obtain clusters without sharp boundaries). Note that values too close to 1 can lead to slow convergence.

Fuzzy clustering with fanny is different from k-means and hierarchical clustering, in that it returns degrees of membership for each observation in each cluster \[15\]. Here we formed three groups, so that we can represent the degree of membership in RGB colors. In comparison with other fuzzy clustering methods, fanny has the following features:

- it also accepts a dissimilarity matrix (also called distance matrix; describes pairwise distinction between objects).
- it is more robust to the spherical cluster assumption (The clusters are expected to be of similar size, so that the assignment to the nearest cluster center is the correct assignment).
- it provides a novel graphical display, the silhouette plot \[16\].

5. Packages in R

R is a language and environment for statistical computing and graphics. R provides a wide variety of statistical (linear and non-linear modelling, classical statistical tests, time-series analysis, classification, clustering, ...) and graphical techniques, and is highly extensible.

One of R’s strengths is the ease with which well-designed publication-quality plots can be produced, including mathematical symbols and formulae where needed. Great care has been taken over the defaults for the minor design choices in graphics, but the user retains full control.

R is an integrated suite of software facilities for data manipulation, calculation and graphical display. It includes:

- an effective data handling and storage facility,
- a suite of operators for calculations on arrays, in particular matrices,
- a large, coherent, integrated collection of intermediate tools for data analysis,
- graphical facilities for data analysis and display either on-screen or on hardcopy, and
- a well-developed, simple and effective programming language which includes conditionals, loops, user-defined recursive functions and input and output facilities.

Many users think of R as a statistics system. We prefer to think of it as an environment within which statistical techniques are implemented. R can be extended (easily) via packages \[17\].

Many packages provide functionality for more than one of the topics listed below, the section headings are mainly meant as quick starting points rather than an ultimate categorization.

Except for packages stats and cluster (which ship with base R and hence are part of every R installation), each package is listed only once \[18\].

The xlsx package gives programmatic control of Excel files using R. A high level API (Application Programming Interface - specifies how some software components should interact with each other) allows the user to read a sheet of an xlsx document into a data.frame and write a data.frame to a file. Lower level functionality permits the direct manipulation of sheets, rows and cells. For example, the user has control to set colors, fonts, data formats, add borders, hide/unhide sheets, add/remove rows, add/remove sheets, etc.

6. Creating RGB maps

Using the Fanny algorithm, we obtained 3 fuzzy clusters. Each of them is identified with one RGB color. Depending on the membership degree to a given cluster, we assign the value of RGB (0-255) to each unit. Each color is formed by mixing of three intensities R, G and B color, respectively (Figure 1). The result is a 1x1 square km which forms a mixture of RGB colors according to the membership degree to the cluster.

![Figure 1. Mixture of RGB colors according to the membership degree.](image)

The green color indicates the best quality — no or smallest effect by acidification. The red color indicates a slight effect on soil by acidification. And the blue one represents relatively the worst damage and acidification in these cases shows extreme values. With the clustering analysis we obtained also 3 clusters, that we call R (red), G (green), and B (blue). (G - the best, R - middle and B - the worst).

Fuzzy cluster analysis (Figure 2) indicated that differentiation of soils occurs only in the upper horizon. Diagnostic horizon has very similar values of chemical indicators in all clusters. Its status can be assessed as good. For all clusters is the median value of $C$ (concentration of basic cations) above 14cmol/kg and value of $E$ (Content of exchangeable alkalies)
is bigger than 50. Somewhat worse, but in a global aspect of average values reaches median $S$ (saturation of alkalies), which is below 40%.

Upper horizons are strongly affected by acidification. Although the global status of all clusters can be described as bad, already there is some differentiation.

The first cluster (R) is created within the clusters average. Median $S$ is 2.5%, which is below the critical value of 10%, as C value 1cmol/kg is below the critical value 2cmol/kg. Only the value of E is favorable, takes a value of 4.

The second cluster (G) is relatively the best. Median $S$ is still below the critical value, but 7% it’s obviously better than the first cluster. Median C is 3cmol/kg and is thus just above the critical value. The value of E is favored 12.

The third cluster (B) is relatively the worst. $S$ becomes the extreme value of 1%. Similarly bad is C with value 0.3 cmol/kg. Neither the value of E is not favorable and is lower than 1.

The map in Figure 3 contains results of the classical cluster analysis. Chemical indicators of diagnostic horizon for individual clusters are very similar to the fuzzy cluster analysis.

Upper horizon of first cluster (R) can be considered as relatively moderate. However, its condition is bad. Median of $S$ (saturation of alkalies) is below the critical value of 10%, C (concentration of basic cations) is with a value of 2.5 cmol/kg just above the critical value of 2 cmol/kg. Only the value of E (Content of exchangeable alkalies) is positive and median has value 10.

All indicators are below the critical values, including the value C. Specific values of medians are $S$ 2%, kvk 0.5 cmol/kg and the value of E is less than the 1.

The upper horizon of third cluster (G) can be described as average acidified, but within our group represents the best condition. Median of $S$ with a value of 15% is above the critical limit of 10%. 7cmol/kg is the global average of value C. The value of E is 26.

7. Future Work

In the future work I would like to determine the damage degree of soil horizons by fuzzy decision making system. In this case I will use a LFLC (Linguistic Fuzzy Logic Controller), program which is based on theory of fuzzy sets and fuzzy logic. I will try to describe a situation of soil horizons by linguistically formu-
lated rules. Results will be showed on maps and compared with results of fuzzy clustering.

8. Conclusion

Soil degradation is a huge worldwide problem. Currently, the concept of soil degradation is used in a much broader sense than a few decades ago. The term relates to all processes that can lead to soil degradation. In our work, we have focused on soil acidification.

The goal of the work is to determine the degree of damage to soil horizons. It is all about long-term damage by soil acidification, which is regarded as the main factor of soil degradation in the Czech Republic. We searched areas which have the same composition or the same soil damage throughout the country.

The reason why we used fuzzy clustering analysis is nondisjunction, the biggest advantage over classic clustering analysis. In this case, we used the fuzzy cluster analysis for creation of RGB map (Fig. 1) and classical cluster analysis (Fig. 2). The maps represent areas of soil horizons influenced by acidification and the degree of damage. In the case of clustering analysis is differentiation of soil acidification only in the upper horizon as in the case of fuzzy clustering.

Chemical indicators of the individual clusters are similar to the fuzzy clustering analysis. The advantage of fuzzy clustering analysis is not only in showing areas damaged by acidification but also in showing more degrees of damage in each square. While the classical analysis shows three levels of damage, using fuzzy analysis, we determine more degrees of damage.

Thanks to fuzzy clustering analysis we can determine to damage degree of soil horizons more precisely. Also we can find areas which have totally different level of damage by acidification. (neighboring squares which have totally different color intensities) These areas are the subject of further detailed research on geology and geography.

Sometimes it is not possible to describe one property of an object accurately, but only vaguely, ambiguously or with some uncertainty. From the description of the properties of this object it is possible to select the value that would do the best. Acidification of soils in the Czech Republic and its consequences for their chemistry mostly work as a predisposition factor. It can unfavourably affect the condition of forest ecosystems in relation to other factors.

Of course, possible destructions of this subsystem can result in a serious reduction of the production level of forest stands or even in the collapse of a forest ecosystem as a whole. Therefore potential consequences for changes and development of soil processes and soil properties are important from the aspect of present evaluation as well as from the aspect of their future development [2].

References