ROMANIAN JOURNAL OF INFORMATION SCIENCE AND TECHNOLOGY
Volume **26**, Number 3–4, 2023, 276–288

Approach to Evaluate the Data of Moss Biomonitoring Studies: Preprocessing and Preliminary Ranking

Gheorghe DUCA[1], Sergey TRAVIN[2], Inga ZINICOVSCAIA[1, 3, 4], and Radu-Emil PRECUP[5, 6, *]

[1]Institute of Chemistry, Research Center of Physical and Inorganic Chemistry, Str. Academiei 3, 2028 Chisinau, Republic of Moldova

[2]Russian Academy of Sciences, Semenov Federal Research Center for Chemical Physics, Kosygina Street 4, Building 1, 119991 Moscow, Russian Federation

[3]Joint Institute for Nuclear Research, Str. Joliot-Curie 6, 141980 Dubna, Russian Federation

[4]Horia Hulubei National Institute for R&D in Physics and Nuclear Engineering, Str. Reactorului 30, MG-6, Bucharest-Magurele, Romania

[5]Politehnica University of Timisoara, Department of Automation and Applied Informatics, Bd. V. Parvan 2, 300223 Timisoara, Romania

[6]Romanian Academy – Timisoara Branch, Center for Fundamental and Advanced Technical Research, Bd. Mihai Viteazu 24, 300223 Timisoara, Romania

E-mails: `ggduca@gmail.com`, `travinso@yandex.ru`, `zinikovskaia@mail.ru`, `radu.precup@aut.upt.ro` [*]

[*] Corresponding author

Abstract. This paper, dedicated to Acad. Florin Gheorghe Filip, at his 15th anniversary as the Chairman of the Information Science and Technology Section of the Romanian Academy, suggests an approach to describe the data obtained in biomonitoring studies using mosses, on the example of the Republic of Moldova. In total, 33 moss samples were collected on the territory of Moldova, the elemental composition of which was determined by neutron activation analysis and atomic absorption spectrometry. At the first stage of the work, a correlation analysis was carried out with the ranking of data in the order of decreasing total correlation, which made it possible to preliminarily reduce the number of elements to two factors. At the stage of data sorting, iron, the content of which in the environment can be associated with anthropogenic activity, was chosen as the element determining the rank. The next stage of work was data smoothing using a discrete cosine transform, for which the codes were rewritten and the algorithm was ported to the Excel-VBA environment, which is most suitable for preprocessing and graphical display of experimental data. Also, an algorithm was developed for determining the number of linearly independent (basis) vectors in which the matrix itself can be decomposed. Two principal roots were identified, of which the larger one is several thousand units, and three lower roots, which are less than unit were excluded, since their absolute values differ by two to three or more orders of magnitude downward from the eigenvalues of the influencing components. The paper is dedicated to Acad. Florin Gheorghe Filip, at his 15th anniversary as the Chairman of the Information Science and Technology Section of the Romanian Academy, and at his 75th anniversary.

Key-words: Algorithm; correlation analysis; data preprocessing; moss biomonitoring.

1. Introduction

The world population is currently over 7.9 billion people. More than 200 thousand are born daily and around 150,000 people die every day around the world. Of these, about 100,000 die from age-related diseases, about 19,000 – from air pollution, an estimated 14,000 people die every day from water pollution and roughly 10,000 deaths per day a caused by COVID-19 as highlighted in [1].

These figures demonstrate that air pollution is one of the environmental health risks in the world, and require elaboration of urgent mission to reduce the levels of pollution and to save millions of lives [2]. Controlling anthropogenic air pollutants is a very complex problem and is currently done through automatic devices, however they provide data on a limited number of pollutants and economic aspects have to be integrated, particularly their high costs prevent the development of dense monitoring networks [2, 3]. The ability of plants to assimilate trace metals from the surrounding atmosphere and bioaccumulate them in their tissues has made plants the most suitable biological air monitors [4].

Passive moss biomonitoring is low-cost monitoring technique widely applied in many European countries [5–8] to asses the spatial distribution of airborne pollutants in ecosystems [9]. Mosses are ideal to evaluate air pollution, because of their diversity of habitats, structural simplicity and rapid multiplication rate. Mosses can accumulate metals above their physiological needs due to the absence of cuticle in their tissues and the abundance of sites with exchangeable cations in their cell wall [10]. Although the heavy metal concentration in mosses provides no direct quantitative measurement of deposition, this information can be derived by using regression approaches relating the results from moss surveys to deposition monitoring data [5].

Nowadays, Factor analysis (FA) has become a principal statistical method of investigation in environmental studies and is widely used method in evaluating the resulting concentrations from biomonitoring using mosses [6–8, 11]. FA is a multivariate tool which allows to determine both the source apportionment and composition of the sources without a priori knowledge of the sources and their composition [12]. It explains the variance of observed variables using a smaller number of potential variables – the factors. The goal is to reduce the number of variables and reveal the structure of the relationship between variables. The weaknesses of FA lie in the ambiguity of the estimation of factor parameters (*i.e.*, the dependence of the FA result on the rotation used) and in the need to specify the number of common factors before performing the analysis [6].

This paper is dedicated to Acad. Florin Gheorghe Filip, at his 15th anniversary as the Chairman of the Information Science and Technology Section of the Romanian Academy, and at his

75th anniversary. The first and the fourth authors of this paper are grateful to Acad. Filip for enabling their contact, which led to the start of a fruitful cooperation.

This paper is built upon authors' recent paper [13] on processing, neural network-based modeling of biomonitoring studies data and validation on Republic of Moldova data, and it proposes an approach for describing the data gained by biomonitoring with the use of mosses on the example of data obtained for the Republic of Moldova in 2015/2016 moss survey. The paper is structured as follows: materials, methods, results and discussion are treated in the next section. Section 3 is focused on data processing. The covariance matrix and the principal component extraction are presented in Section 4. The conclusions are drawn in Section 5.

2. Materials, Methods, Results and Discussion

During the Spring of 2015, 33 samples of moss Hypnum cupressiforme were collected on the entire territory of the Republic of Moldova. The studied areas as well as the procedure of samples collection and preparation for chemical analysis are described in [14]. A part of the results of the research conducted in this paper is summarized in [15].

Two analytical techniques, neutron activation analysis and atomic absorption spectrometry, were applied to determine the content of 42 elements (Na, Mg, Al, Cl, K, Ca, Sc, Ti, V, Cr, Mn, Fe, Co, Ni, Cu, Zn, As, Se, Br, Rb, Sr, Zr, Cd, Sb, Cs, Ba, La, Ce, Nd, Sm, Eu, Gd, Tb, Tm, Yb, Hf, Ta, W, Pb, Th and U) in moss samples. The detailed description of the performed analysis is given in [14].

The content of determined elements in moss samples is presented in Table 1 given in [15]. According to the approach proposed in the present study, it is very important to compare the distribution of elements with a priori measured parameters of the model, for example, with geographic coordinates, the level of industrial energy consumption, etc., that were not taken into account when applying FA. In the present study, a table of 34 rows (settlements) and 34 columns (the first two columns are geographic latitude and longitude) was taken for analysis, the remainder are the content of 32 elements in mosses. It should be emphasized that the squareness of the table is not required. The coincidence of the number of geographic points with the number of elements supplemented by two columns of geographic coordinates should be regarded as occasional. In general, the number of rows can be either more or less than the number of columns, but it cannot be less than the number of linearly independent components that define the rank of the covariance matrix of the model.

The first stage in the work was the correlation matrix compiling, which is necessary for two reasons: both for a visual representation of what is "what looks like" and for ranking the elements according to the degree of compatibility. Correlation analysis is a statistical method of data processing by which density of correlation ratio between two or more (32 ones in this paper) variables is measured [16].

In the correlation matrix, the correlation coefficient close to zero means that the values, most probably, are unrelated. In practice, values less than 0.5 in modulus are no longer subject to consideration and are usually not tried to be reduced to a single formula. Table 2a given in [15] shows the correlation coefficients between the studied elements. For visual appearance, conditional formatting was applied to the table using a color scale: the closer to 1.0 the correlation coefficient, the stronger the green tone dominates in the color. Uncorrelated cells are colored with a predominant use of red tones. Finally, intermediate values are colored yellowish. However, in this form, as it will be shown in the next section, the correlation matrix is relatively uninformative. It becomes much more apparent if the elements are ranked in the order of decreasing of total correlation as it will be shown in the next section. In this case, the sum of the correlation coefficients for the row is used as a variable for sorting the rows. Also, the elements with high values of the correlation coefficient are accumulated closer to the upper left corner of the table, while the elements with low values of the correlation coefficient remain in the lower right corner.

According to Table 2b given in [15], a maximum of two factors are required to describe the initial data. The first will allow describing the group behavior of an association of 20 elements from chromium to sodium (in the order used after ranking the correlations, *i.e.* Cr, Ta, Co, Th, Tb, Fe, Sc, La, Rb, As, Cs, Yb, Ce, Eu, Sm, Ni, U, Hf, Tm, Na). The second group, possibly, will consist of an association of four elements Ti, Al, Mg, V, which do not have high correlation with chromium, but confidently correlate with each other with a coefficient no less than 0.96. Seven elements Ba, Sb, Ca, K, Mn, Zn, Sr, Br have very low correlation coefficients with other groups of elements, which did not allow distinguish them into separate factor. It should be mentioned that in the entire table, only a couple of cells have negative values. They are highlighted in purple and in fact their values are more likely to be zero than negative.

3. Data Preprocessing

According to [17] Box and Wilson (1992), when working with a dataset, dimensioning of the original data is required, *i.e.* reduction them to a comparable scale. In our case, the concentration of various elements at different collection sites varies from hundredths to thousands of μg/g, *i.e.* by five orders of magnitude. For this reason, direct comparison of data on concentrations is meaningless. If to follow directly the recommendations of Box, then for each of the 32 determined elements it would be necessary to calculate the average value for all geographical locations and set it as zero, relative to which variability of data up and down would be reduced using a linear (proportional) transformation to the range [–1, 1].

Such approach, being mathematically flawless, still raises serious criticism from a physicochemical point of view. First, for further decoupling of the basis vectors of pollution, in order to avoid the so-called rotational ambiguity is very useful to draw on considerations of nonnegativity of concentrations.

In this sense, it would be preferable to reduce all data to the range [0, 1]. But this is not a solution either, since the loss of obviousness of more representative elements in comparison with "small impurities" is not compensated with the achieved simplicity.

For the analysis of pollution in the studied area, the most reasonable would be the transition from the concentrations to their decimal logarithms. In this case, the variation for each element will practically be reduced to a single range, since the concentrations scattering practically in none of the cases does not exceed an order of magnitude. However, the hierarchy of concentration precedence will also remain, since the logarithm is a monotonic increasing function.

Table 3 given in [15] shows the ranking in decreasing order of values of the correlation coefficients between the logarithms of the measured concentrations. The comparison with Table 2b given in [15] shows that the order changes insignificantly. The elements that had tendency to associate in one group remained in their groups. But after switching to a logarithmic scale, the changes (displayed by the color range) became smoother and the color change became softer.

The next step of data compilation is an attempt to sort them not by correlation coefficients, but by primary values of concentrations. As preliminary data processing showed, neither the alignment of points in latitude, nor in longitude, nor in distance from any geographic or administrative center does not bring any order to the rather chaotic spread of concentration values. Consequently, we have renounced to this approach.

It would be ideal to plot the data in descending order of pollution. However, that sequence of lines (the order of geographical locations) that leads to a decrease in the content of one element, will not necessarily lead to a decrease in the content of another. Nevertheless, most of the data have high, close to 1.0, correlation coefficients with the leader (either chromium in conventional analysis or tantalum in logarithmic analysis). All correlations for the logarithms of concentrations, without exceptions, are positive. Therefore, without loss of generality the list can be ranked by almost any element. Iron was chosen as the element determining rank, since iron content in the environment may be associated with anthropogenic activity. In addition, unlike tantalum or thorium, iron is not a trace element and is at the beginning of the list of prominently represented elements, second only to the main soil elements Ca, K and Al, and significantly exceeding Na, Mg, and other elements. Further in the present study, the data will be used, in which the order of the human settlements – sampling points were ranged by the iron content.

The next step in data preparation for statistical processing is ranking of elements by representativeness. For this purpose, the matrix, in the rows of which there are elements and their concentrations, and the leading column contains the names of settlements sorted "by iron" is transposed. Now the leading column contains the names of the elements, and the final column contains the integral representation (the sum of the logarithms of concentrations) of the elements throughout the territory. After sorting of the elements according to the integral contamination calculated in this way, the matrix acquires the form in which in the rows of the leading column geographical locations and the elements themselves are listed in the first row. In all cells of the table, the logarithms of the measured concentrations of elements will be given.

Assuming that the content of elements is strongly correlated, it is not supposed to have upward or downward outliers in the rank distribution. They should be more or less smooth and certainly monotonous, for which the use of anti-aliasing seems to be useful.

One of the efficient and stable smoothing procedures is described in detail in [18]. That procedure is based on *discrete cosine transform* (DCT), it provides reliable smoothing of equidistant data in one or more dimensions.

The underlying "penalized" approach to least squares method will be briefly discussed as follows, in the specific case of successful application to smoothing of high-frequency noise of a spectroscopic signal. This approach consists in minimizing an objective function (or criterion) that balances the fidelity to the data, measured by the *residual sum-of-squares* (RSS), and a *penalty term* (P) that reflects the roughness of the smooth data in terms of seeking to minimize the objective function

$$F(\hat{\mathbf{y}}) = RSS + s \cdot P(\hat{\mathbf{y}}) = ||\hat{\mathbf{y}} - \mathbf{y}||^2 + s \cdot P(\hat{\mathbf{y}}). \tag{1}$$

The notation $||\hat{\mathbf{y}} - \mathbf{y}||^2$ in (1) indicates the squared Euclidean norm, namely the sum of the squares of the deviations, and the penalty term gives the degree of signal coarsening based on the divided differences of the second order (generalization of the concept of derivatives to the discrete case):

$$P(\hat{\mathbf{y}}) = ||\mathbf{D}\hat{\mathbf{y}}||^2, \tag{2}$$

where $\mathbf{D}$ is a tridiagonal square matrix, that assumes the simplest form for the equidistant data (the rank goes exactly with a unit step):

$$\mathbf{D} = \begin{bmatrix} -1 & 1 & 0 & \dots & 0 \\ 1 & -2 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & 0 \\ 0 & \dots & 1 & -2 & 1 \\ 0 & \dots & 0 & 1 & -1 \end{bmatrix}. \tag{3}$$

The decomposition of the matrix $\mathbf{D}$ in terms of eigenvectors gives the representation

$$\begin{aligned} &\mathbf{D} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{-1}, \\ &\mathbf{\Lambda} = diag(\lambda_1, \dots, \lambda_n), \\ &\lambda_i = -2 + 2\, cos[(i-1)\pi/n], i = 1 \dots n, \end{aligned} \tag{4}$$

where $\lambda_i, i = 1 \dots n$ are the eigenvalues. For a smoothed signal, the following expression is applied:

$$\begin{aligned} &\hat{\mathbf{y}} = \mathbf{U} \cdot (\mathbf{I}_n + s \cdot \mathbf{\Lambda}^2)^{-1} \cdot \mathbf{U}^T \cdot \mathbf{y} = \mathbf{U} \cdot \mathbf{\Gamma} \cdot \mathbf{U}^T \cdot \mathbf{y}, \\ &\mathbf{\Gamma} = diag(\gamma_1, \dots, \gamma_n), \\ &\gamma_i = \{1 + s[2 - 2\, cos[(i-1)\pi/n]]^2\}^{-1}, i = 1 \dots n, \end{aligned} \tag{5}$$

and $\mathbf{I}_n$ is the n^{th} order identity matrix.

Further, the DCT and the *inverse cosine transform* (IDCT) are applied. DCT is similar to the Fast Fourier Transform, but ignores its imaginary (sine) part. The formula to compute the DCT is

$$\begin{aligned} &t(u,v) = c(u)c(v) \sum_{k=0}^{N-1} \sum_{l=0}^{M-1} I(k,l)\, cos[(2k+1)u\pi/(2n)]\, cos[(2l+1)v\pi/(2m)], \\ &i = 0 \dots N-1, c(i) = \begin{cases} \sqrt{1/N} & \text{if } i = 0, \\ \sqrt{2/N} & \text{if } i \neq 0, \end{cases} \end{aligned} \tag{6}$$

and the formula to compute the IDCT is

$$I(k,l) = \sum_{u=0}^{N-1} \sum_{v=0}^{M-1} c(u)c(v)t(u,v)\, cos[(2k+1)u\pi/(2n)]\, cos[(2l+1)v\pi/(2m)]. \tag{7}$$

The meaning of the sequential application of the DCT and the IDCT is to select the optimal parameter s, which sets the cutoff of the frequency response of the filter $\mathbf{\Gamma}$:

$$\hat{\mathbf{y}} = \mathbf{U} \cdot \mathbf{\Gamma} \cdot DCT(\mathbf{y}) = IDCT(\mathbf{\Gamma} \cdot DCT(\mathbf{y})). \tag{8}$$

The codes were rewritten and the algorithm ported to the Excel-VBA environment, which is used in preprocessing and graphical presentation of experimental data. It makes sense to set the parameter s with a logarithmic step. Fig. 1 given in [15] demonstrates how the frequency response of $\mathbf{\Gamma}$ changes when it is varied in the range of five orders of magnitude.

In its turn, changing the frequency response of the filter affects the degree of data smoothing. Fig. 2 given in [15] shows a set of partially smoothed profiles for K, the element occupying the 5th position from the end position in terms of the degree of correlation with other elements. At high negative value (p = –5), there is no smoothing and the approximating curve passes through each experimental point. At high positive value (p = 5), as an approximating array an almost horizontal straight line passing through the center of the original array is obtained. Smoothing was reduced to total averaging, that led to a significant loss of information.

Also, an auxiliary program was written and built in order to optimize the smoothing parameter by finding the maximum of the cross-validation function as the *generalized cross-validation* (GCV) expressed as

$$GCV(s) = n\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 / \Big[n - \sum_{i=1}^{n}(1 + s\lambda_i^2)^{-1}\Big]^2 \tag{9}$$

In this version, smoothing occurs completely automatically and does not require initial setting of parameters, and the result corresponds not only to the best from the point of view of visual comprehension smoothing, but to mathematically correct filtering out of noise and experimental errors without losing meaningful information. It is clear that for different elements the primary degree of profile smoothness is different. The more to the right is the element in the cross-correlation table, the worse it is. Fig. 3 given in [15] shows the spline for Sr, occupying the last place in the table.

Smoothing of th elements from the middle of the table is is easily acceptable, since there is less initial data variability. Fig. 4 given in [15] shows smoothing for Cr, an element in the middle of the correlation table. In this case, p = 1.5, *i.e.* 20 times less coercion was applied to the primary data than in the case of Sr.

Fig. 5 given in [15] shows a smoothed curve for Fe, which differs in the minimum way from the primary one and the smoothing parameter $s = 10^p$ is still 100 times less than for Cr.

It can be seen that the element Fe was justifiably chosen for the ranking. Although Ta and Th seem to show even higher adherence to associativity with other elements, they they are significantly less representative in concentration, which overshadowing them, on the set of signs they are of lesser importance than iron.

4. Covariance Matrix and Principal Component Extraction

After a matrix of 1088 elements was compiled (34 rows of sampling points and 32 columns of the content of chemical elements, not counting two geographic coordinates), the next step is to determine the number of linearly independent (basis) vectors into which the matrix can be decomposed.

From the point of view of linear algebra, the problem is formulated as follows. The matrix of experimental data **D** can be represented as the product of two smaller matrices **C** and $\mathbf{S}^T$. In this case, the conventionally vertical matrix (represented as a set of columns) **C** is a representation of the "concentration" profiles for each simulated component in the system. The word "concentration" is put in quotations because in fact it is referred to a set of values of the factors Fi for each collection site. The height of this matrix is qual to the total number of collection sited, in our case 34. The number of columns is a priori unknown, but this paper aims to minimize it in order to represent the decomposition of the data matrix **D** in terms of the lowest dimensional in terms of

$$\mathbf{D} = \mathbf{C}\mathbf{S}^T + \mathbf{E}, \tag{10}$$

where **E** is an error matrix, that means a residual variation in the dataset that is not associated with any chemical contribution.

It needs to be emphasized once again that the number of columns in the concentration matrix **C** corresponds to the dimension of the basis of independent variables. At the same time, $\mathbf{S}^T$ is a transposed matrix of "spectra" of the corresponding polluting factor. In the matrix **S**, the rows are the spectra of individual (or expected) pollutants, the number of which is equal to the rank (the number of linearly independent components of the matrix **D**), and the columns, if considering them as spectra, are the readings in the corresponding spectroscopic channel. In our case, the role of the spectrum is played by the content of each of the elements, and the number of the spectroscopic channel is the number of the chemical element (namely, the ordinal number in the correlation table).

The matrix of obtained data $\mathbf{D} \in \Re^{N_R \times N_C}$ is assumed to be information redundant, *i.e.* the number of rows in it N_R (the number of collection sites) is obviously greater than the number of independent components in a chemical mixture, and the number of columns N_C is the number of individual chemical elements that are similar to spectroscopic channels, *i.e.* allow to distinguish one spectrum from another outside the noise by the unique concentration ratio of 32 elements included in the analysis. The concentration ratio of elements is the "fingerprint" or chemical spectrum of pollution.

Actually, the whole problem is the most reasonable decoupling of **D** into factors **C** and **S**, and to do this blindly when a priori information about the reference spectra and concentrations is not available. It is clear that such a solution is ambiguous, but the information obtained during the decomposition is valuable in itself. The problem is formulated in terms of the *Multivariate Curve Resolution Alternating Least Squares* (MCR-ALS).

A very similar in appearance (up to alphabetical symbol) and mathematically completely equivalent statement of the problem is used in the method of *independent component analysis* (ICA) characterized by

$$\mathbf{X} = \mathbf{A}\mathbf{S}^T + \mathbf{E}, \tag{11}$$

where $\mathbf{A} \in \Re^{I \times N}$ is the so-called mixing matrix in the notation specific to ICA and also the concentration matrix in the notation specific to MCR ALS. The number I of rows in **A** is also the number of sampling sites, and the number N of columns is the dimension of the basis of independent vectors. The matrix $\mathbf{S}^T \in \Re^{N \times J}$ is the transposed matrix of reference spectra, where N is the number of standards, and **J** is the number of determined chemical elements-marker. The error matrix, which also appears in (10), is $\mathbf{E} \in \Re^{I \times J}$.

The integers I and J, are also the number of rows and columns, respectively, of the data matrix **X**. The integer N is also the number of components included in the bilinear form of the expansion of (11). ICA algorithms try to find the "decoupling" matrix **W**, which is the inverse of the mixing matrix **A** according to

$$\hat{\mathbf{S}}^T = \mathbf{W}\mathbf{X}, \tag{12}$$

where $\hat{\mathbf{S}}^T$ is the characterization of the parent matrix $\mathbf{S}^T$.

According to (12), when $\mathbf{W} = \mathbf{A}^{+}$, *i.e.* $\mathbf{W}$ is the pseudo-inverse of the mixing matrix $\mathbf{A}$, the estimated source signal $\hat{\mathbf{S}}^T$ will be equal to the initial (usually called original) source signal $\mathbf{S}_T$:

$$\hat{\mathbf{S}}^T = \mathbf{W}\mathbf{X} = \mathbf{W}\mathbf{A}\mathbf{S}^T = \mathbf{S}^T. \tag{13}$$

Although the described algorithms have been already implemented for many times, for example, in Matlab and in specialized literature there are numerous references to distributives, most of these links do not allow an external user to get real access to the running software. In addition, all published algorithms use iterative procedures for determining the decoupling matrix, therefore the scope of the present study is to develop the own package for blind processing of spectroscopic sets, focused on using the most efficient calculation methods, so that iterative procedures can be avoided wherever possible, in particular to refuse to basic use of fittings in the least squares method.

This paper produces original techniques suitable for solving some problems of spectroscopic recognition and separation. If at the first stage to leave without consideration the problem of rotational and massive ambiguity of the definition of the basis vectors, then obtaining of their "working set" is reduced to solving the problem of the eigenvalues of the covariance matrix as performed by Danilevsky's method [19]. The essence of Danilevsky's method is to reduce the matrix $\mathbf{A}$ using $n - 1$ similar transformations to a similar Frobenius matrix $\mathbf{P}$:

$$P = \begin{bmatrix} p_1 & p_2 & p_3 & \cdots & p_{n-1} & p_n \\ 1 & 0 & 0 & \ldots & 0 & 0 \\ 0 & 1 & 0 & \ldots & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & 1 & 0 \end{bmatrix}. \tag{14}$$

In the Frobenius form, only the elements of the first row and units below the main diagonal are present in the matrix. All other elements are equal to zero. This form is ideal for generating characteristic polynomials and determining eigenvalues. The coefficients of the characteristic polynomial are obtained as follows by expanding the determinant of the matrix in the first row:

$$\begin{aligned} &det(\mathbf{P} - \lambda\mathbf{I}_n) = (p_1 - \lambda)(-\lambda)^{n-1} - p_2(-\lambda)^{n-2} + \cdots + (-1)^{n-1}p_n \\ &= (-\lambda)^n + p_1(-\lambda)^{n-1} - p_2(-\lambda)^{n-2} + \cdots + (-1)^{n-1}p_n = 0. \end{aligned} \tag{15}$$

Danilevsky's algorithm is guaranteed to give a set of coefficients of the characteristic polynomial. Even in the case of disperse matrices with singularities, when the "leading" element of the row becomes zero, then the characteristic polynomial splits into the product of two characteristic polynomials, each of which can be obtained by the same algorithm. Since Danilevsky's method is described with sufficient details in [19], this paper will not dwell on it in more detail.

An important place is occupied by the choice of an automated software algorithm for finding the roots of the characteristic polynomial. The overwhelming majority of known algorithms for finding roots are suitable for general function (not only for polynomials), but they have two hard-to-remove drawbacks, (1) and (2): (1) the necessity of choosing a starting approximation for a root and setting it manually, that is incompatible with the concept of "automated procedure"; (2) uncertain behavior of the algorithm in the case of multiple roots, since this zero slope of the function graph appears in this case, which excludes the use of tangent or secant methods. Among other things, in (2), the function can have the same sign to the left and right of the root, which excludes dichotomy.

An interesting algorithm for calculating the real roots of polynomials was developed by I. Kostin and posted for general use at [21]. The main idea of this algorithm is quite simple and can be summarized in two sentences. First, the real root of the polynomial is always located in the area of monotonic variation of the polynomial, *i.e.* between the roots of the derivative of the polynomial. Second, in turn, the derivative of the polynomial is also a polynomial, however of a lesser degree, and, having found its real roots, it is necessary to search for the roots of the original polynomial between the roots of the derivative by dividing the segment in half.

It should be noted the excellent level of comments within the text of the program posted at [20] and the impeccable quality of its work. As far as this paper is concerned, over several years of intensive use, both for practical purposes and for specially organized "difficult cases" with multiple roots, very widely spaced groups of roots of high degree polynomials, the authors have not found a single program failure.

Having received a set of eigenvalues, it is then necessary to discriminate between them. This is illustrated in Table 3 given in [15].

The point is that the method of principal components is based on the fact that the contribution of the basis vector to the variation is the greater, the greater its eigenvalue. Therefore, the lower roots, as a rule, are responsible not for significant changes in the spectrum, but for random noises. In any case, they do not make a significant contribution to the explained part of the dependencies. For this reason, they should be excluded from the scheme of further analysis, as illustrated in Fig. 6 given in [15]. In the case treated in this paper, it would seem appropriate to leave the two principal roots, of which the larger one is several thousand units, and exclude the three lowest roots, which are less than 1.0.

The resulting set of roots of the characteristic polynomial coincides with the number of linearly independent basis vectors, only for ideal spectra without noise. In practice, both for experimental and model sets, the eigenvalues of the lowest absolute value give eigenvectors that only formally represent the "basic standar", but in fact do not contain any information except noise. But such spectra, firstly, are easily determined visually, and secondly, insignificant eigenvalues can be deleted even programmatically, since their absolute values differ by two to three or more orders of magnitude from the eigenvalues of the really influencing components.

The least squares method was applied in the previous section. However, metaheuristic algorithms are also popular recently as, for example, quantum annealers [21], monarch butterfly optimization algorithms [22], slime mould algorithms [23], particle swarm optimization algorithms [24], and string theory algorithms [25]. These algorithms are efficient, however the mathematical guarantee of their convergence is generally not solved.

5. Conclusions

This paper developed an approach to describe the biomonitoring data obtained for moss samples collected in the Republic of Moldova. At the first stage of the work, using correlation analysis it was possible to reduce the number of determined chemical elements to two factors. Iron proved to be suitable element for data ranking. The data smoothing was performed using a discrete cosine transform, for which the codes were rewritten and the algorithm was ported to the Excel-VBA environment.

An algorithm for determining the number of linearly independent (basis) vectors in which the matrix itself can be decomposed was developed. Two principal roots were identified, of which the larger one is several thousand units, and three lower roots were excluded from further discussion.

Future research will be focused on the continuation of the description of the data processing procedure. Being relatively easily understandable, the approach will be applied to data measured in other representative applications including well-being [26], evolving systems [27] and telesurgical applications [28].

Acknowledgement. This work was supported by the grant of the Romanian Ministry of Education and Research, CNCS - UEFISCDI, project number PN-III-P4-ID-PCE-2020-0269, within PNCDI III, supported by the national project of Republic of Moldova number 20.80009.5007.27.

References

[1] Official Coronavirus Cases. Accessed: July 17, 2023. [Online]. Available: https://www.worldometers.info/.

[2] S. GIORDANO, V. SPAGNUOLO and F. CAPOZZI, *Biomonitoring of air pollution*, Atmosphere **12**(4), 2021, paper 433.

[3] S. CHAKRABORTTY and G. T. PARATKAR, *Biomonitoring of trace element air pollution using mosses*, Aerosol and Air Quality Research **6**(3), 2006, pp. 247–258.

[4] T. MANCHENO, R. ZALAKEVICIUTE, M. GONZÁLEZ-RODRÍGUEZ and K. ALEXANDRINO, *Assessment of metals in PM10 filters and Araucaria heterophylla needles in two areas of Quito, Ecuador*, Heliyon **7**(1), 2021, paper e05966.

[5] H. HARMENS, D. A. NORRIS, E. STEINNES, E. KUBIN, J. PIISPANEN, R. ALBER, Y. ALEKSIAYENAK, O. BLUM, M. COŞKUN, M. DAM, L. DE TEMMERMAN, J. A. FERNÁNDEZ, M. FROLOVA, M. FRONTASYEVA, L. GONZÁLEZ-MIQUEO, K. GRODZIŃSKA, Z. JERAN, S. KORZEKWA, M. KRMAR, K. KVIETKUS, S. LEBLOND, S. LIIV, S. H. MAGNÚSSON, B. MAŇKOVSKÍA, R. PESCH, A. RÜHLING, J. M. SANTAMARIA, W. SCHRÖDER, Z. SPIRIC, I. SUCHARA, L. THNI, V. URUMOV, L. YURUKOVA and H. G. ZECHMEISTER, *Mosses as biomonitors of atmospheric heavy metal deposition: Spatial patterns and temporal trends in Europe*, Environmental Pollution **158**(10), 2010, pp. 3144–3156.

[6] A. S. KRAKOVSKÁ, V. SVOZILÍK, I. ZINICOVSCAIA, K. VERGEL and P. JANČÍK, *Analysis of spatial data from moss biomonitoring in Czech-Polish border*, Atmosphere **11**(11), 2020, paper 1237.

[7] O. CHALIGAVA, S. SHETEKAURI, W. M. BADAWY, M. V. FRONTASYEVA, I. ZINICOVSCAIA, T. SHETEKAURI, A. KVLIVIDZE, K. VERGEL and N. YUSHIN, *Characterization of trace elements in atmospheric deposition studied by moss biomonitoring in Georgia*, Archives of Environmental Contamination and Toxicology **80**, 2021, pp. 350–367.

[8] I. ZINICOVSCAIA, C. HRAMCO, O. CHALIGAVA, N. YUSHIN, D. GROZDOV, K. VERGEL and G. DUCA, *Accumulation of potentially toxic elements in mosses collected in the Republic of Moldova*, Plants **10**(3), 2021, paper 471.

[9] O. MOTYKA, I. PAVLKOV, J. BITTA, M. FRONTASYEVA and P. JANČÍK, *Moss biomonitoring and air pollution modelling on a regional scale: delayed reflection of industrial pollution in moss in a heavily polluted region?*, Environmental Science and Pollution Research **27**, 2020, pp. 32569–32578.

[10] G. MACEDO-MIRANDA, P. AVILA-PÉREZ, P. GIL-VARGAS, G. ZARAZÚA, J. C. SÁNCHEZ-MEZA, C. ZEPEDA-GÓMEZ and S. TEJEDA, *Accumulation of heavy metals in mosses: a biomonitoring study*, SpringerPlus **5**, 2016, paper 715.

[11] A. S. KAPLUNOVSKY, *Factor analysis in environmental studies*, HAIT Journal of Science and Engineering B **2**(1–2), 2005, pp. 54–94.

[12] Y.-L. XIE, P. K. HOPKE, P. PAATERO, L. A. BARRIE and S.-M. LI, *Identification of source nature and seasonal variations of Arctic aerosol by the multilinear engine*, Atmospheric Environment **33**(16), 1999, pp. 2549–2562.

[13] R.-E. PRECUP, G. DUCA, S. TRAVIN and I. ZINICOVSCAIA, *Processing, neural network-based modeling of biomonitoring studies data and validation on Republic of Moldova data*, Proceedings of the Romanian Academy, Series A: Mathematics, Physics, Technical Sciences, Information Science **23**(4), 2022, pp. 403–410.

[14] I. ZINICOVSCAIA, C. HRAMCO, O. G. DULIU, K. VERGEL, O. A. CULICOV, M. V. FRONTASYEVA and G. DUCA, *Air pollution study in the Republic of Moldova using moss biomonitoring technique*, Bulletin of Environmental Contamination and Toxicology **98**, 2017, pp. 262–269.

[15] G. DUCA, S. TRAVIN, I. ZINICOVSCAIA and R.-E. PRECUP, *Supplementary material of the paper Gheorghe Duca, Sergey Travin, Inga Zinicovscaia and R.-E. Precup, "Approach to Evaluate the Data of Moss Biomonitoring Studies: Preprocessing and Preliminary Ranking"*, Romanian Journal of Information Science and Technology, 2023. Accessed: Jul. 25, 2023. [Online]. Available: http://www.aut.upt.ro/∼rprecup/Supplementary_material_ROMJIST2.doc.

[16] T. J. ARCHDEACON, *Correlation and Regression Analysis: A Historian's Guide*, University of Wisconsin Press, Madison, WI, 1994.

[17] G. E. P. BOX and K. B. WILSON, *On the experimental attainment of optimum conditions*, Journal of the Royal Statistical Society Series B **13**, 1951, pp. 1–45.

[18] D. GARCIA, *Robust smoothing of gridded data in one and higher dimensions with missing values*, Computational Statistics & Data Analysis **54**(4), 2010, pp. 1167–1178.

[19] G. R. KOZIN, *Algorithms for Numerical Methods of Linear Algebra and Their Software Implementation* (in Russian), National Research Nuclear University MEPhI (Moscow Engineering Physics Institute), Moscow, 2019.

[20] I. KOSTIN, *Algorithm for calculating the real roots of polynomials* (in Russian). Accessed: Jul. 25, 2023. [Online]. Available: https://habr.com/ru/post/303342/.

[21] S. V. ROMERO, E., OSABA, E. VILLAR-RODRIGUEZ, I. OREGI and Y. BAN, *Hybrid approach for solving real-world bin packing problem instances using quantum annealers*, arXiv preprint, arXiv:2303.01977, 2023.

[22] Y.-H. FENG, S. DEB, G.-G. WANG and A. HOSSEIN ALAVI, *Monarch butterfly optimization: A comprehensive review*, Expert Systems with Applications **168**, 2021, paper 114418.

[23] R.-E. PRECUP, R.-C. DAVID, R.-C. ROMAN, A.-I. SZEDLAK-STINEAN and E. M. PETRIU, *Optimal tuning of interval type-2 fuzzy controllers for nonlinear servo systems using slime mould algorithm*, International Journal of Systems Science DOI: 10.1080/00207721.2021.1927236, 2021.

[24] Z. C. JOHANYÁK, *A modified particle swarm optimization algorithm for the optimization of a fuzzy classification subsystem in a series hybrid electric vehicle*, Technicki Vjesnik – Technical Gazette **24**(2), 2017, pp. 295–301.

[25] L. RODRÍGUEZ, O. CASTILLO, M. GARCÍA VALDEZ and J. SORIA, A new meta-heuristic optimization algorithm based on a paradigm from physics: string theory, Journal of Intelligent & Fuzzy Systems **41**(1), 2021, pp. 1657–1675.

[26] F. G. FILIP, *Automation and computers and their contribution to human well-being and resilience*, Studies in Informatics and Control **30**(4), 2021, pp. 5–18.

[27] S. BLAŽIČ, D., DOVŽAN and I. ŠKRJANC, Cloud-based identification of an evolving system with supervisory mechanisms, Proceedings of 2014 IEEE International Symposium on Intelligent Control, Antibes, France, 2014, pp. 1906–1911.

[28] R.-E. PRECUP, T. HAIDEGGER, S. PREITL, B. BENYÓ, A. S. PAUL and L. KOVÁCS, Fuzzy control solution for telesurgical applications, Applied and Computational Mathematics **11**(3), 2012, pp. 378–397.