Annales Societatis Geologorum Poloniae (2006), vol. 76: 215–224.

NUMERICAL CORRELATION OF MANY MULTIDIMENSIONAL GEOLOGICAL RECORDS

Adam WALANUS[1] & Dorota NALEPKA[2]

[1] *Institute of Technology, Faculty of Mathematics and Natural Sciences, Rzeszów University, ul. Rejtana 16 A, 35-959 Rzeszów, Poland; e-mail: walanus@univ.rzeszow.pl*
[2] *W. Szafer Institute of Botany, Polish Academy of Sciences, ul. Lubicz 46, 31-512 Kraków, Poland; e-mail: nalepka@ib-pan.krakow.pl*

Walanus, A. & Nalepka, D., 2006. Numerical correlation of many multidimensional geological records. *Annales Societatis Geologorum Poloniae*, 76: 215–224.

Abstract: It is frequent task to correlate profiles or cores basing on different measurements performed on the series of samples. The difficulty arises when there are many profiles and none is the main or reference one. The reason is that the number of possible correlations grows exponentially with the number of profiles. To resolve the problem a Monte Carlo method is adopted here, what makes it very probable to discover the best correlations in a reasonable amount of computing time. The quality of a correlation is measured by a metric of dissimilarity of the samples. The final result, given in graphical form, has a form of lines connecting correlative samples from different profiles. The number of lines (correlations across profiles) is user-defined and can vary from one to dozens. The number of profiles, samples, and variables depends only on the computational resources. Large problems need longer computation times to achieve stable results.

Key words: Monte Carlo, computer intensive, dissimilarity coefficient, data standardization.

Manuscript received 18 Ocoter 2005, accepted 18 May 2006

INTRODUCTION

Probably, the most frequently occurring piece of numerical data in geology is a record of measurements preformed along (as a rule vertical) profile, or core (well-log). Since typically many features (variables) are investigated, the record is multidimensional. Having two or more "parallel", neighbouring profiles, with the same measurements performed, it is natural to correlate them. While there are many methods of correlation of two profiles (Birks, 1986), the more difficult is to correlate many (three, tens) of profiles. Such a task can be reduced to many correlations of two profiles, when one of profiles can be treated as the main one. However, if there were no geological reason for treating one profile as reference, such solution would introduce a subjective bias into the resulting correlation. Moreover, sequential correlation of profiles with the reference one neglect the mutual information connected with each pair of profiles. If there are NP profiles, there are NP-1 correlations with the reference profile, while there are as many as NP*(NP-1)/2 correlations of different pairs of profiles.

In fact, if there is array of many profiles, parallel in sense of importance, to be mutually correlated, the numerical method used have to mirror the geological situation, i.e. to find the general correlation, which is the best one for all the profiles, at the same time.

Computational difficulties arising with the fast increase of amount of possible correlations with the number of profiles (NP) are overcome by the use of approximate Monte Carlo method. While the random method, in case of very large problems (many long profiles with many features), can not assure that the found solution is really the best one, the correlation probably will be close to optimal. For small and medium size problems (depending on computer resources), finding of the best correlation is very probable.

The described below algorithm is implemented in program MultCorr (see Fig. 1) (Nalepka, 2005).

STRUCTURE OF DATA FOR ANALYSES

A basic data unit here is a spreadsheet or a table with NV variables and NL levels or samples taken from a single profile. The variables (measurements of different features of sampled material), as a rule, are ordered in columns; the samples are ordered in rows. For the computer application described here, the first column should contain depths of samples, and the first row variable names. Identifiers (names) of variables must be consistent in all the correlated profiles. Some variables may be absent for some profiles; the order of variables in the tables does not matter. The levels are ordered stratigraphically, what is natural.

Consistent variable names enable definition of "a set of variables" to be used in calculations. Trying different variables in correlating profiles (the question of "feature selection"; Guyon & Elisseeff, 2003) seems to be crucial in many fields of applications. Variables are assumed to be quantitative or almost quantitative; however, there is no strict constraint in that point.

THE PROBLEM

There is a number (NP) of records (tables, profiles) to be correlated. All profiles, in principle, should contain correlative horizons, i.e. samples to be found as similar. If one or a few profiles are completely different than most of others, the result will be skewed.

There is no limit for the number of profiles to be correlated, other than memory resources for storing them. Of course, the computing time is increasing with increasing NP. However, even for large NP (dozens) provisional calculations can be fast. To achieve more precise and stable results, longer computing times would be necessary. Generally, the precision of the results seems to increase logarithmically with the number of trials (nT – user defined main parameter, roughly proportional to the computing time).

The sense of correlation or synchronization of profiles, expressed graphically, that is in appropriate levels, in all profiles, are to be connected by lines (Fig. 1). Lines connecting the most similar samples, one from each profile, at the same time divides all the profiles. In the following text, such line is referred to as a division. After the first division is found, the next one can be searched for. The obvious constraint is that lines connecting samples can not cross one another. They can have common samples (the lines can touch), but all samples from one division must be older or younger than those from the other division (except for possibility of common samples). Number of divisions (ND) is a user-defined parameter; it can be set from 1 to 100.

In the simplest approach, the first connection joins the most similar samples; the second one is less optimal, and so on. Such an algorithm would work quite fast. However, searching, in consecutive steps, for increasingly unimportant division is not a good method, because it does not reflect reality.

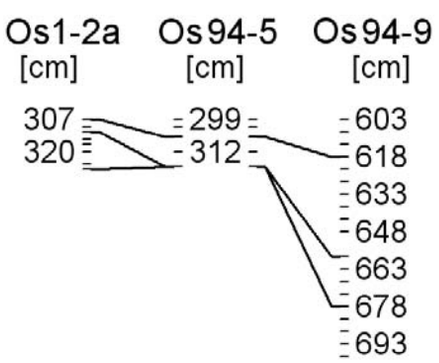

Fig. 1. Examplary correlation of three (NP = 3) very short profiles. The number of divisions ND = 3. Lines of correlations can "use" one sample twice (here, the bottommost sample from Os 94-5) or many times; however, they cannot cross one another

Crucial for the algorithm described below is an assumption that there is no natural hierarchy of divisions.

The main idea is to divide profiles in a synchronized way by connecting the most similar samples, where "the most similar" refers to the total measure of similarity of samples within groups (ND groups of NP samples in each group). In the case of sequential algorithms, previous divisions would block off newer ones, because natural restriction does not allow divisions to cross one another. Thus, the resulting total similarity would not be maximal.

CRITERION OF GOODNESS OF CORRELATION

For any pair of levels (samples), the dissimilarity coefficient (DC) can be calculated (Gower & Legendre, 1986; Maher, 1998). The simplest form of DC is the sum, over all variables included, of absolute values of difference between the values from the first and the second profile. It is the so-called Manhattan metric (Maher, 1998), because it resembles distance from one point to the other to be walked in the rectangular net of streets. Such a definition is adopted in the algorithm; however, with possible application of different data transformations. The variety of possible other definitions of DC (e.g., Euclidian – square root of sum of squares) will not be discussed here, because they are computationally irrelevant to the main concept of algorithm.

It is necessary to note that scaling of variables is important for DC, if distances for different variables are to be summed up. For example, for one variable distance between samples is measured in [gm^{-3}], because the variable refers to density, while for the other variable, grain size, it is measured in [mm]. Numerical addition of values measured in different units must be carefully performed. If one variable has values of the order of 100 and the other 0.01, then the influence of the second variable on the resulting DC value will be completely negligible. The simple way to manage such situations is to standardize variables (see below, *Transformation of variables*).

The quality of correlation of NP profiles by ND lines (divisions) is measured by the value of total DC. The elemental DC is calculated for the pair of samples. The total DC is the sum of DCs for all pairs of samples, provided that both samples in each pair came from the same division. For one division there are NP*(NP-1)/2 pairs of samples, so the number of involved DCs is ND* NP*(NP-1)/2.

Taking into account the number of variables (NV) used in calculation of the elemental DC, the number NV*ND*NP*(NP-1)/2 of differences (the most deeply elemental DC) is involved. This number is used for normalization of the total DC, to make it comparable among different analyses (under the assumption that variables were normalized or that they are of similar nature).

THE ALGORITHM FOR SEARCHING THE LOWEST TOTAL DC

Number of possible correlations

In case of one division only (ND = 1), and profiles of, say 100 samples each, the total number of possible divisions is 100^{NP-1}. For five profiles (NP = 5), this number is 10^8. Since in the calculation of DC for one (trial) correlation about NV*(NP-1) subtractions are involved, the time of computation needed to check all possible correlations would be of the order of an hour (assuming NV = 100, and typical 3GHz PC).

For more divisions (ND>1), the number of operations rises very fast with ND, easily approaching non-realistic computing time. The solution then would be the application of the Monte Carlo method (e.g., Robert & Casella, 1999). However, for not very low NP, ND, and number of samples, simple Monte Carlo trials (of randomly chosen correlations) can, in a realistic timeframe, check only little percentage of all possibilities. This is why some compromise has been adopted in the present algorithm.

The algorithm

Starting description from the most deeply nested pieces of the algorithm, the following operations are performed.

(1) All of the NP profiles are numbered randomly. It is assumed that no one profile is a reference one, and all the profiles are of equal importance. In the following text, the notion "first profile", "second", and so on, refers to the random order.

(2) From the first profile, the sample is randomly selected from those not yet used in any previously performed division. The uniform probability distribution is used, so all samples have equal probability of being selected.

(3) If it is not the first division, one has to recognize to which section of the profile the selected sample belongs. Since profiles are assumed to be in the stratigraphical order, divisions can not cross one another. Samples from the subsequent profiles will be considered only from this, appropriate section.

(4) In the second profile (or in its fragment), the sample is searched for the lowest DC with the sample already selected in the first profile. Either all possible samples, or only a given number of randomly selected samples are being checked. That last option, application of which is up to the user, is for limiting computation time in case of large problems.

(5) Repeat point (4) for the next profiles. However, starting from the third profile, for calculation of elemental DC, instead of simple values of variables, the average values calculated for already connected samples are used and compared with simple values from the actual profile. At this step, it is possible to apply a "penalty" for overly close divisions. The value of DC is multiplied by a factor $(1+\text{penalty}/(1+\text{distance})^{1/2})$, where distance is the number of samples between divisions in the considered profile, and Penalty is a user-defined parameter.

(6) The obtained DC is checked if it is lower than the formerly obtained lowest value. If it is lower, then its value is remembered, as well as the related division, for use in the higher-level steps of the algorithm.

(i) According to the idea of Monte Carlo trials, the steps (1)–(6) are repeated many times. Let, that number of trials in the nested loop is denoted by nT_1. The lowest obtained DC is recorded, as well as the related (optimal) division.

(ii) The point (i) is repeated ND times to complete all divisions required by the user. The total (at this stage) DC is calculated. The calculation is as follows: within each division, elemental DCs for all pairs of samples from different profiles are calculated (NP*(NP-1)/2 pairs) and summed up. This is repeated for all ND divisions, and in total DC all elemental DCs are summed up. In the calculation of the total DC, all profiles and all divisions are of equal statistical weight.

(iii) The point (ii) is repeated many times (nT_2 – number of trials in the outer loop). The lowest obtained total DC is recorded, as well as the related correlation. This correlation (comprising ND divisions) is the final one.

Parameters of the algorithm

The main parameter of the algorithm it is the number of Monte Carlo trials (nT, or *Number of Trials* in the program interface). Computing time increases linearly with the number of trials. A practical option in the program is to declare a time limit. The trials are stopped when the time limit is reached.

The user-defined nT is then recalculated into two numbers mentioned in the previous paragraph: nT_1 and nT_2. The ratio of nT_2/nT_1 is also user-defined (*Main/Sub-trials*). While a larger *Number of Trials* give higher precision, the second parameter affects the calculations less clearly. Its value (default = 1) can be adjusted experimentally. However, it seems to be important only for large data sets (see the discussion below, in *Exemplary results of correlation – Artificial data*). Generally, higher values of *Main/Sub-trials* increase the probability of obtaining the best correlation, but also increase the number of trials (i.e. computing time); whereas lower values assure good results in a reasonable amount of time even for very large data sets.

The next parameter has a goal similar to the previous one: limiting computing time without degradation of reliability. Instead of checking the DC value for all samples (in step (4) of the algorithm), only a few randomly selected samples are checked. The number of samples to be checked is set by the user as a value of the parameter *Try samples* (the default value is *All*).

In case of multiple divisions (ND>1) it is possible to enforce avoidance of overly close divisions. If, in every profile there are small fragments similar to each other, then all divisions (sample connections) can (optimally) be indicated in those fragments only, although other fragments might be interesting as well. The parameter *Penalty for too close* can help in such a case (Fig. 2).

The last parameter, *Number of divisions* (ND), unlike the previously described ones, is "visible" in the result. It is the number of divisions of profiles, or number of connections between samples from different profiles. It is worth

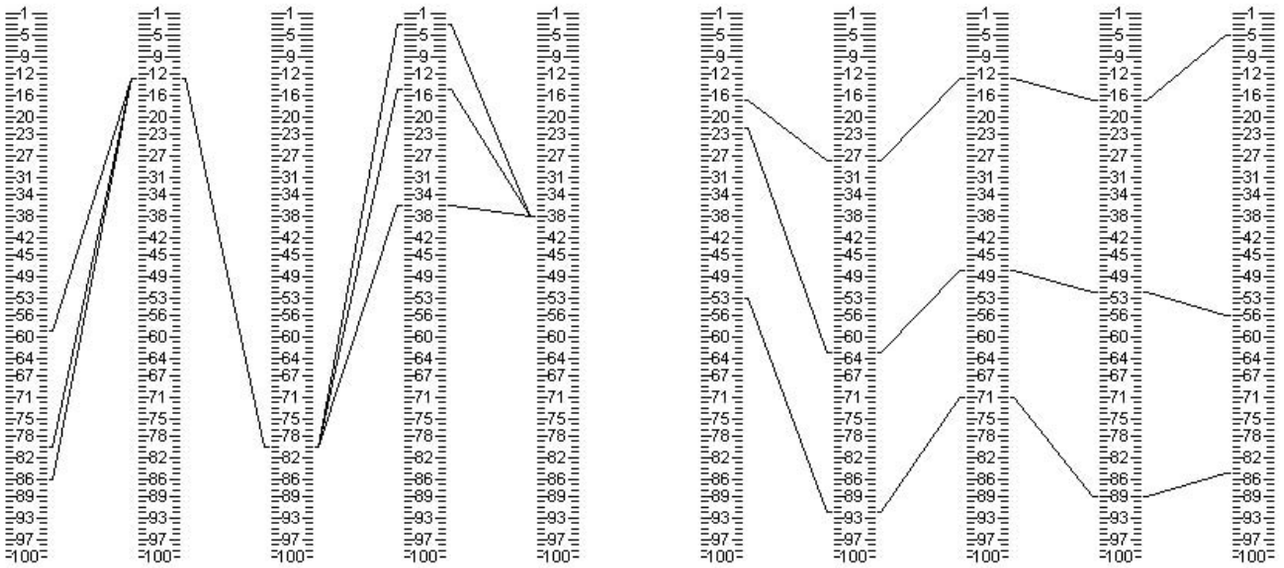

Fig. 2. The effect of the application of parameter *Penalty for too close*. The left graph – zero penalty, the right graph – penalty equals 0.5. The left graph does not illustrate the optimal correlation since in case of zero *Penalty* all three divisions should be identical

mentioning again that the algorithm operates in such a way that there is no order of importance in divisions. As a result we get ND divisions, which are the best in general (at least close to the best, since the algorithm is not deterministic).

TRANSFORMATION OF VARIABLES

Variables are, as a rule, to be standardized, since the dissimilarity coefficient (DC) is calculated over many variables. Standardization is not necessary, and if no applied variables of higher variability (standard deviation) will simply weight more in the analysis. Of course, in case of differences as high as an order of magnitude, the smaller range variables would have almost no influence on the calculated DCs.

Other type transformation of data can be applied to achieve some special effects. For example, square root transformation diminishes relatively the influence of high values, what may be desirable from some point of view.

Variables standardization (two kinds)

Standardization of a variable consists of recalculation of its values by subtracting the average value, and dividing the result by the standard deviation of that variable. The resulting standardized variable has zero mean and unit standard deviation. Such a transformation operates well for normally distributed variables, or variables not too far from normality. In geology, variables are frequently positively skewed, and could be "normalized" by taking logarithm (if, in place of possible zero values, a reasonable detection limit can be used – what is impossible, for example, in counting individuals of some kind).

There is a certain number of profiles (NP) in the analysis, and each profile has a certain number of samples.

Standardization can be performed separately, within individual profiles, or globally, as single standardization for all values (of a given variable) from all profiles. The first approach is better if, for example, one profile has generally lower values in some variable. Since it would be impossible to find similar values in other profiles, in such a case, it would be better to standardize each profile to the same zero mean and unit standard deviation. However, important information can be lost in such a transformation. A profile with generally low values may be actually synchronous with part of other profile, which in other parts has high values. When standardization is made separately for such individual profiles, finding proper correlation can be difficult, if not impossible. That is the reason for the second kind of standardization, over all profiles.

Global standardization (right graph) changes nothing but the order of magnitude of values, which is sensible only in comparison with other variables.

Global standardization of variables, along all the profiles, does not change the relations between profiles (compare the left and the right graph in Fig. 3). The only reason for such transformation of the variables is to make them intercomparable, which is important in DC calculation. If it is assumed that the variables from Fig. 3 have, in both profiles, correct, representative values, it means that the three bottommost samples from the first are similar to the three uppermost samples from the second. In such a case, no transformation or global standardization should be used. However, if values in the first profile are only accidentally lower (for example, because of incorrect measurement calibration), then separate standardization within the profiles can help (middle graph in Fig. 3).

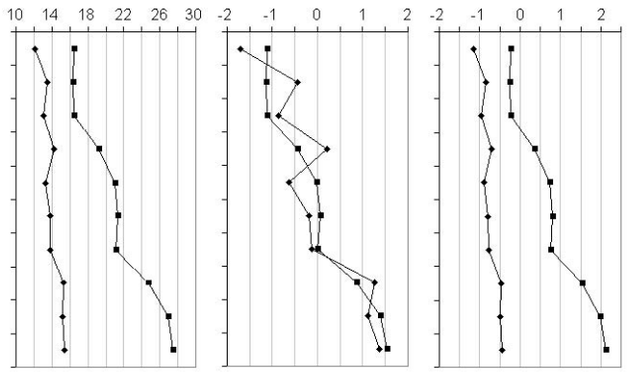

Fig. 3. The effect of the variable standardization. The values of one variable from two profiles are plotted. The original values (left graph) in one profile are smaller than in the other. After independent standardization within profiles (middle graph), the profiles became very similar. Global standardization (right graph) changes nothing but the order of magnitude of values, which is sensible only in comparison with other variables

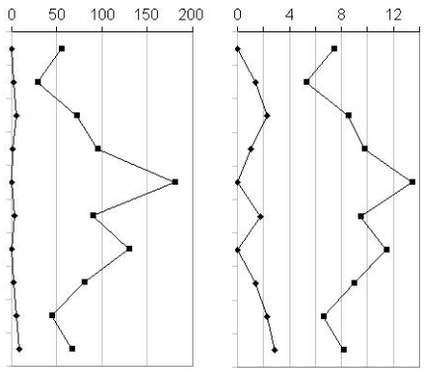

Fig. 4. In the right graph is presented the effect of the *Sqrt* transformation of values (two variables) from the left graph. After transformation statistical weights of the variables with low and high values become comparable

Square root transformation

A square root transformation is especially applicable to counts of some individuals, like pollen grains or other kind of remains. A zero value has a special meaning in such a case (lack of evidence), and cannot be replaced by an arbitrarily low value (as in the log transformation). On the other hand, while some variables have a few counts, some others can have thousands. The generally small amount of individuals in some variable (taxon), by no means indicates a low significance of this variable. However, simple standardization is seldom applied in such cases (Birks & Gordon, 1985). Sqrt (Square root) transformation is typical (Prentice, 1980).

The Sqrt transformation (Fig. 4) influences both the inner variable relations (between samples) and the relations between variables. The first effect is a by-product (which does not skew the final result); the second is the main goal of the transformation. The main idea of *Sqrt* can be exemplified numerically: while the two differences 9-0 = 9, and 9-1 = 8, differ only by 12%, after data transformation they are 3-0= 3 and 3-1= 2, and differ by 40%.

Selection of variables

Selection of variables to be used by the algorithm is essential for the final result. However, use of one variable instead of another is not in the strict sense a variable transformation; it can be treated as a kind of transformation of data for analysis.

In the computer program which implements the described algorithm, no method for automatic variable (feature) selection (Guyon & Elisseeff, 2003) is proposed, since there is no single clear criterion of "a good fit" in the problem of profiles correlation. In fact, selection of variables (as well as data transformations) can be recommended as a tool to obtain an interpretable result. A priori information about variables can not be ignored. The charge of subjectivity in analysis can not be avoided in any non-trivial problem. On the other hand, it is impossible to obtain any desired correlation by manipulating the variables. Application of numerical algorithm imposes a significant amount of objectivity onto the analysis.

Smoothing window

For geological profiles autocorrelation of samples is typical. It means that the neighbouring samples are, as a rule, similar. That fact can be useful for correlation of profiles, especially in case of noisy data, i.e. if the "signal" to be used in analysis is hidden, to some extent, by the noise of different origin (as in. e.g., Fig. 5). The noise, or at least some part of it, is not auto-correlated; averaging neighbouring samples can improve the "signal to noise" ratio.

The parameter *Smoothing window* is the number of neighbouring samples to be added to the one actually considered (in DC calculation). In fact, the weighted average is calculated, with the "triangular" weights. For example, for *Smoothing window* = 1 the weights are: 1/4, 1/2, 1/4, for *Smoothing window* = 2, the weights are: 0.111, 0.222, 0.333, 0.222, 0.111. The highest weight is given to the actually considered sample. The effect of smoothing for the correlation is illustrated in Fig. 6.

EXEMPLARY RESULTS OF CORRELATION

Artificial data

Let five records (NP = 5) of 100 samples each (Fig. 7) consist of NV = 30 variables (the last, of course not visible in the figure). The data are random numbers from a uniform distribution in the interval (0, 1). As a result, no "true" correlation exists in those records.

The result of correlation given in Fig. 7 is obtained for modified data. Simply, the sample Nr 10 from the first record, the sample 20 in the second record, 30 in the third, 40 in the fourth, and 50 in the fifth record are set identical. As a

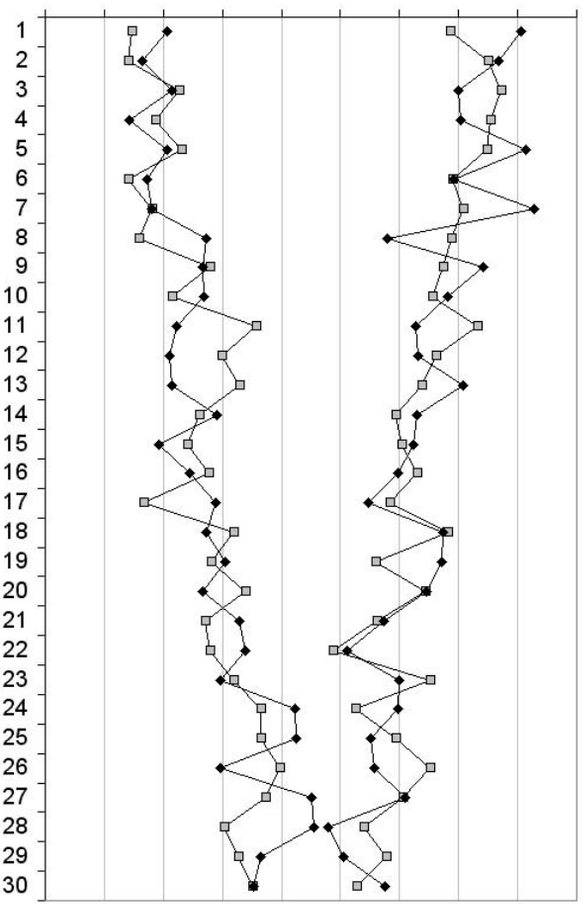

Fig. 5. Exemplary data for presentation of the *Smoothing window* option. There are two profiles (indicated by different point signatures), with two variables, nearly monotonously changing along the profiles. In case of both variables, the values in both profiles are similar along the profiles, but the presence of noise make precise correlation difficult

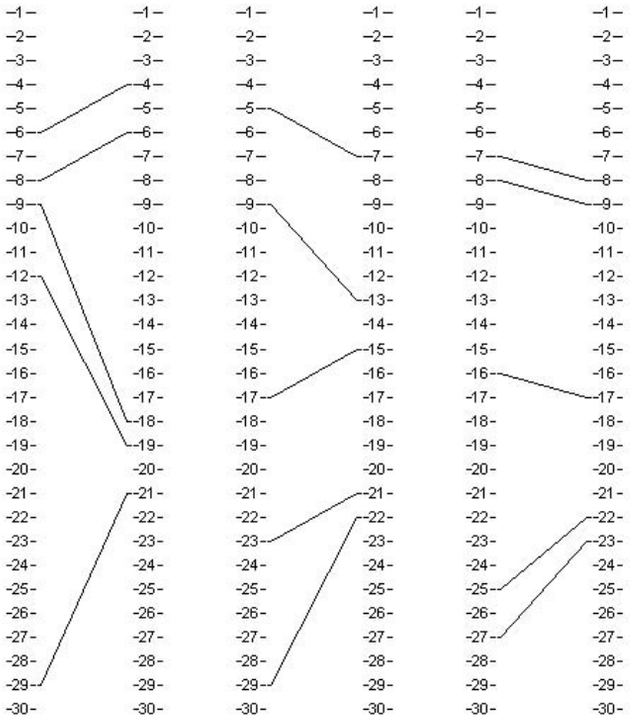

Fig. 6. Correlation of the records from Fig. 5 (ND = 5). The parameter *Smoothing window* is set equal to 0, 1, and 3, from the left plot to the right one. In the ideal correlation lines of divisions are expected horizontal

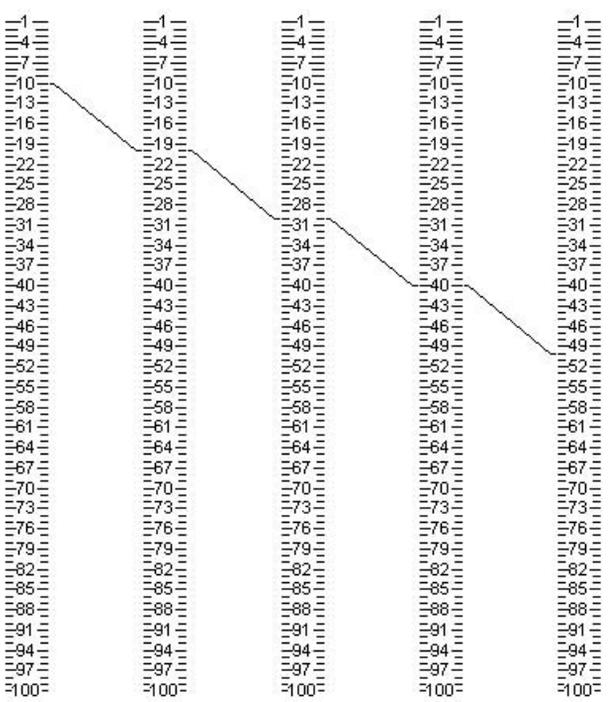

Fig. 7. Artificial example based on random profiles with one sample common for all profiles. Correct correlation is presented

result, very strong "true" correlation exists in the data. The described algorithm can search for such a correlation. Using "illegal" a priori information that there is one level in each record similar to some level from other records, the number of divisions is chosen to be ND = 1. Correct correlation is obtained in computing time of about a second for 1,000 trials. In as low a number as 100 trials, half of the obtained results indicate the proper correlation.

More interesting is correlation of records of strictly random data, with no correlation. Anyway, there are more similar samples among records and also the most similar ones (with the lowest DC).

Using such data the influence of the parameter nT_2/nT_1 (*Main/Sub-trials*) has been investigated. Sub-trials should help when large problems are to be solved in reasonable time. However, in case of medium-size problem and long computing time, use of too many sub-trials is a danger (Fig. 8). If there are very many sub-trials, it can happen that in every main trial the same "best" correlation will be found. As a result, the repetition of main trials would be fruitless waste of computing time. In case of more than one division (ND>1), the best first correlation can exclude the best second one, because of possible crossing. The main trials (nT_2) are necessary if the best set of ND correlations is to be found.

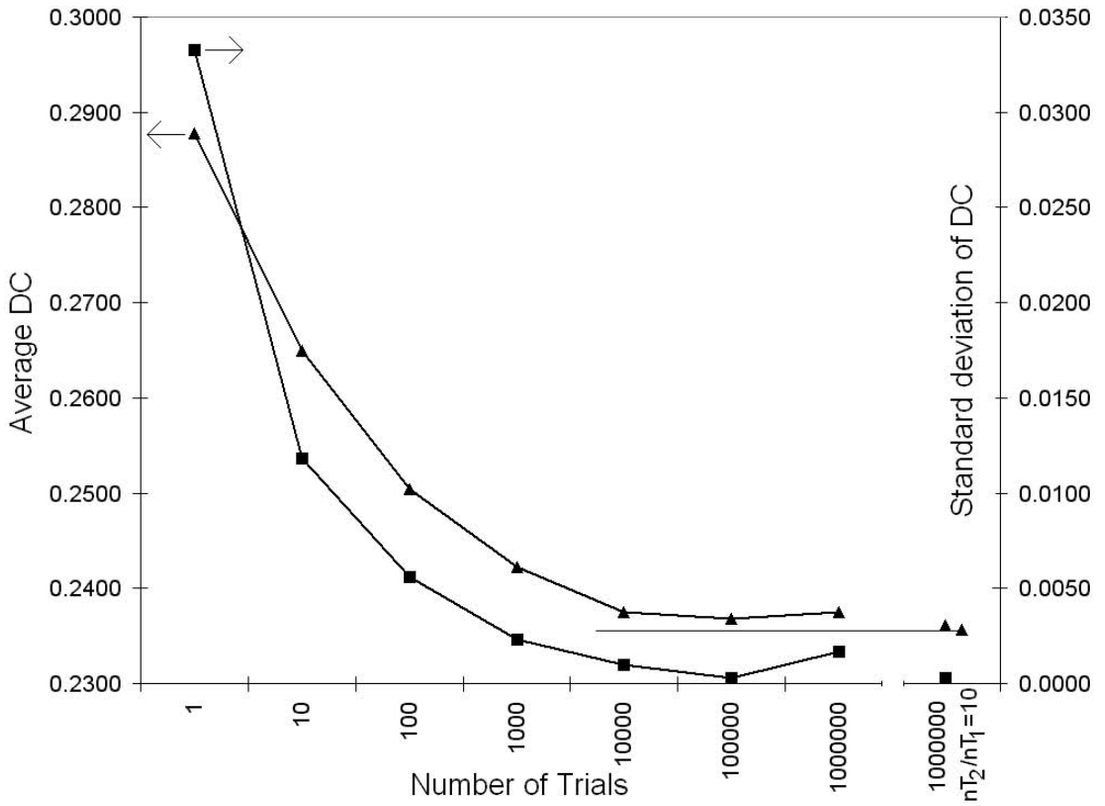

Fig. 8. Illustration of the danger when too many sub-trials (too low value of parameter nT_2/nT_1 = *Main/Sub-trials*) are applied. For large number of trials (here million) the final DC value may be not the lowest one. The general dependence of average DC and its stability (standard deviation) on the number of trials is shown. The parameters were: ND = 3, nT_2/nT_1 = 1, NP = 5, $NL_{1\text{-}5}$ = 100, NV= 30, and uniformly distributed random data were used. The result for nT_2/nT_1 = 10 is given for comparison; beside the average value also the best one is shown (the slightly lower triangle)

Simple exemplary data (Hawaii)

Typical approach to the numerical data analysis contains a kind of comparison of three entities: row data, subjective opinion on the geological situation, and the statistical result itself. Since the subjective element is very important, it is impossible to give really good example, because it should be based on the reader data, what is impossible.

The data used here (Wessel, 2003) consist of only two variables, what makes it easy to visualize data *in extenso* (Fig. 9). In case of one variable the correlation is trivial. However, also in the case of two variables, and seven records, it is clear that correlating records basing on the row data (Fig. 9) is almost impossible.

Data are rather smooth, i.e. not noisy, what results in clear correlation (Fig. 10). Even the application of relatively high *Penalty for too close* would not influence the analysis enough to indicate more correlative levels. However, other kind of standardization of variables results in systematically moved correlation (Fig. 11).

Real data (Quaternary plant pollen counts)

The pollen counts, in the Holocene and the Late Glacial palynological analysis, are integer numbers ranging from 0 to hundreds or more. Typical pollen tables have 100 rows (samples) and 100 columns (variables, pollen taxa). Typical for that kind of data is that some taxa (pine, birch) are abundant, while some others (lime, wheat), by no means less important, are poorly represented by a few pollen grains only. So the square root transformation is applicable here.

Four profiles from Central Poland were used in the analysis: Lake Gopło (Jankowska, 1980), Lake Gościąż (Ralska-Jasiewiczowa *et al.*, 1989), Osłonki (Nalepka, 2005), and Lake Steklin (Noryśkiewicz, 1982). From the taxa present in the pollen tables (almost 300 in case of Lake Gościąż), the number of NV = 21 is used (*Artemisia*, *Betula nana*-t., *Betula*, *Carpinus betulus*, Cerealia undiff., Chenopodiaceae, *Corylus avellana*, *Fraxinus excelsior*, *Hippophaë rhamnoides*, *Juniperus communis*, *Larix*, *Pinus cembra*-t., *Pinus sylvestris*, *Pteridium aquilinum*, *Quercus*, *Rumex acetosa*/*acetosella*, *Salix polaris*-t., *Selaginella selaginoides*, *Tilia* undiff., *Ulmus*, *Urtica* undiff.). The choice of taxa is in principle based on the *a priori* ecological knowledge, and on the goal to be achieved (Holocene or Late Glacial is to be correlated). However, modification of set of taxa after obtaining initial results seems not to be in contradiction with the ideal of objectivity of numerical analysis.

As customary in palynological analysis, data were transformed into percentages within sample – pollen spectrum. Since the percentage calculation is not trivial here (the

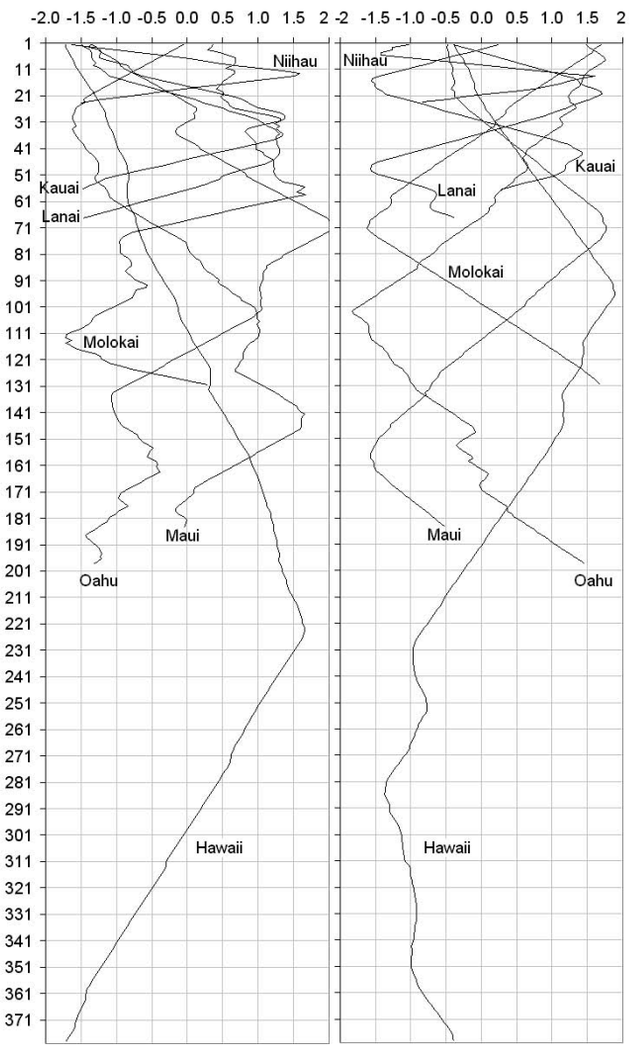

Fig. 9. Exemplary data (available in www, Wessel 2003). Seven records of different number of samples from the region of Hawaii. Two variables, which data consist of, are presented in separate plots. Data are standardized (within records) since variables differ by almost one order of magnitude

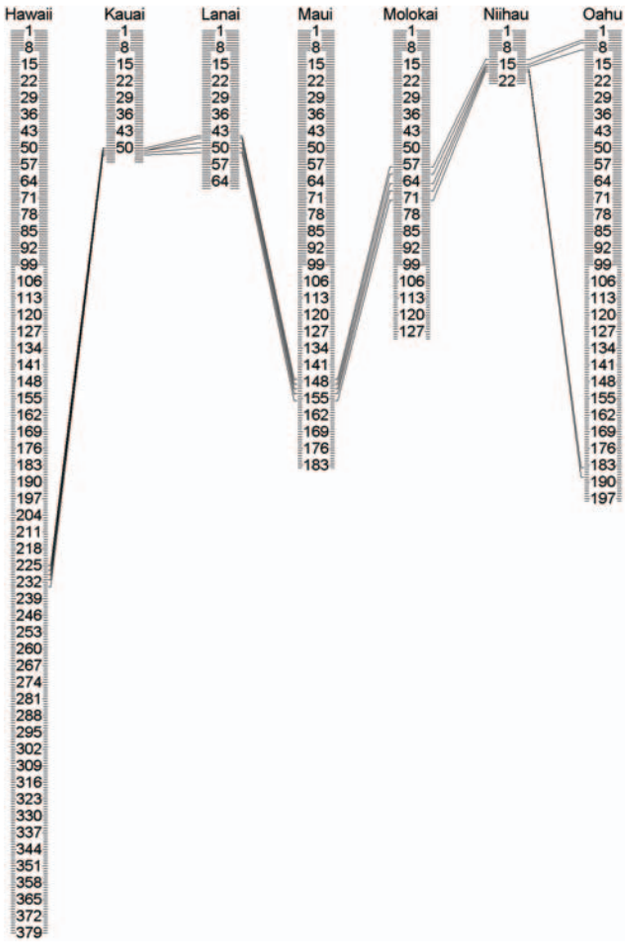

Fig. 10. The result of correlation of records presented in Fig. 9. The global data standardization has been applied; i.e. for each variable the global mean and standard deviation, calculated for all 1,037 samples, has been used for variable standardization

question of base for 100%), it was performed using the POLPAL program (Nalepka & Walanus, 2003; Walanus & Nalepka, 2004), dedicated to pollen counts handling.

The result of correlation obtained for as many as ND= 15 divisions (Fig. 12) is clear. Evident is the synchronous fragment in the lower part of profiles (Late Glacial). The upper parts were probably under deeper local influence, and appear not so similar. However, a correction of the taxa set used in correlation could help to correlate the Holocene part, as long as respective sections are present in all profiles (cf. Nalepka, 2005).

REFERENCES

Birks, H. J. B., 1986. Numerical zonation, comparison and correlation of Quaternary pollen-stratigraphical data. In: Berglund, B. E. (ed.), *Handbook of Holocene Palaeoecology and Palaeohydrology*. Wiley & Sons Ltd., Chichester-New York: 743–774.

Birks, H. J. B. & Gordon, A. D., 1985. *Numerical Methods in Quaternary Pollen Analysis*. Academic Press, London, 317 pp.

Gower, J. C. & Legendre, P., 1986. Metric and Euclidian properties of dissimilarity coefficients. *Journal of Classification*, 3: 5–48.

Guyon, I. & Elisseeff, A., 2003. An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, 3: 1157–1182.

Maher, L. J., Jr., 1998. Slotdeep v. 1.8 adds DC profiles to its DC map. *INQUA Commission for the Study of the Holocene, Working Group on Data-Handling Methods Newsletter*, 18: 4.

Nalepka, D. & Walanus, A., 2003. Data processing in pollen analysis. *Acta Palaeobotanica*, 43 (1): 125–134.

Nalepka, D., 2005. Late Glacial and Holocene palaeoecological conditions and changes of vegetation cover under early farming activity in the south Kujawy region (central Poland). *Acta Palaeobotanica, Suppl*., 6: 3–90.

Noryśkiewicz, B., 1987. Lake Steklin – a reference site for the Dobrzyń-Chełmno Lake District, N. Poland. Report on palaeoecological studies for the IGCP-Project No. 158B. *Acta Palaeobotanica*, 22 (1): 65–83.

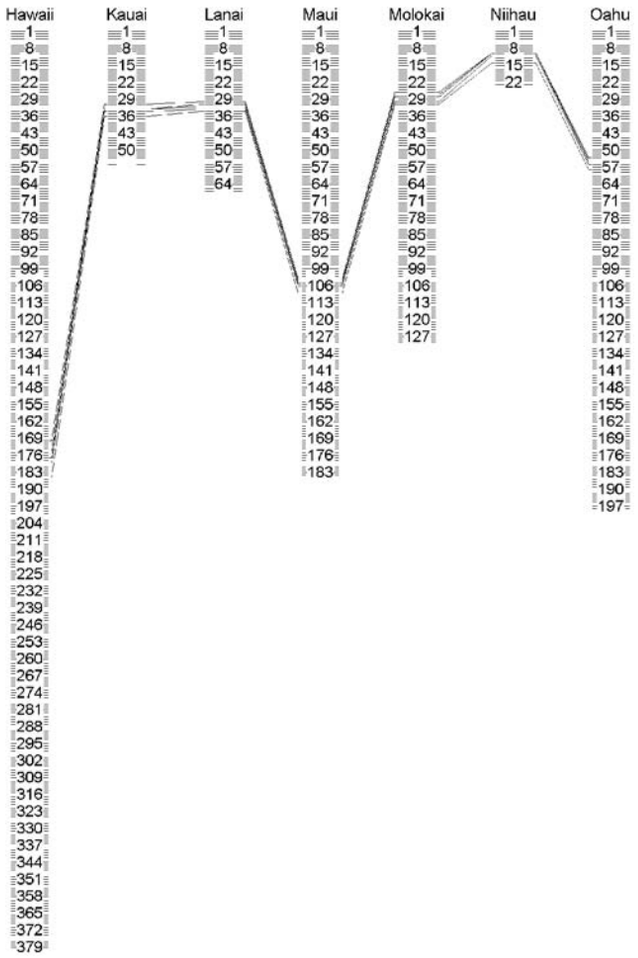

Fig. 11. The result of correlation of records presented in Fig. 9, however, standardized within records. The result clearly differs from that from Fig. 10. If data are treated as precisely measured, than this result should be treated as closer to the true correlation

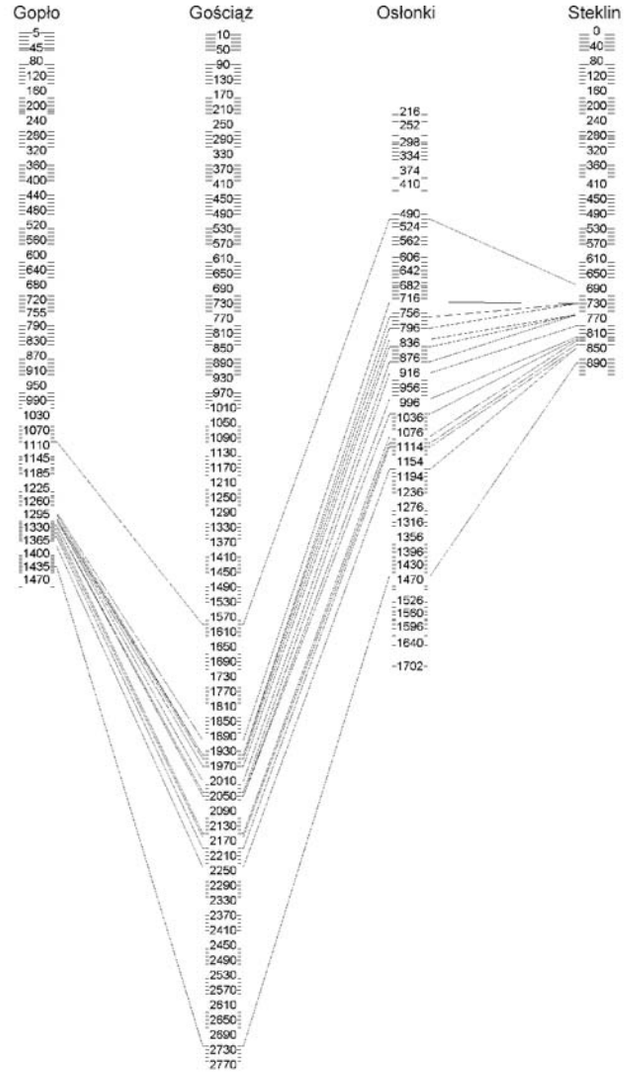

Fig. 12. The result of correlation of four palynological profiles from central Poland. Parameters of analysis are visible in the application "window" in Fig. 13

Prentice, I. C., 1980. Multidimensional scaling as a research tool in Quaternary palynology: a review of theory and methods. *Review of Palaeobotany and Palynology*, 31: 71–104.

Ralska-Jasiewiczowa, M. & van Geel, B., 1998. Human impact on the vegetation of the Lake Gościąż surroundings in prehistoric and early-historic times. In: Ralska-Jasiewiczowa, M., Goslar, T., Madeyska, T. & Starkel, L. (eds), *Lake Gościąż, Central Poland. A Monographic Study. Part 1.* W. Szafer Institute of Botany, Polish Academy of Sciences, Kraków: 267–293.

Ralska-Jasiewiczowa, M., Demske, D. & van Geel, B., 1998. Late-Glacial vegetation history recorded in the Lake Gościąż sediments. In: Ralska-Jasiewiczowa, M., Goslar, T., Madeyska, T. 7 Starkel, L. (eds), *Lake Gościąż, Central Poland. A Monographic Study. Part 1.* W. Szafer Institute of Botany, Polish Academy of Sciences, Kraków: 128–143.

Robert, C. P. & Casella, G., 1999. *Monte Carlo Statistical Methods*. Springer, 536 pp.

Walanus, A. & Nalepka, D., 2004. Integration of Late Glacial and Holocene pollen data from Poland. *Annales Societatis Geologorum Poloniae*, 74: 285–294.

Wessel, P., 2003. http://www.soest.hawaii.edu/wessel/courses/gg313.html, Geological Data Analysis. The School of Ocean and Earth Science and Technology, University of Hawaii.

Streszczenie

KORELACJA NUMERYCZNA WIELOWYMIAROWYCH DANYCH DLA KILKU PROFILI GELOGICZNYCH

Adam Walanus & Dorota Nalepka

Korelowanie dwóch lub kilku sekwencji próbek z profilu, na podstawie wyników różnych pomiarów wykonywanych dla próbek, jest jednym z najczęściej wykonywanych zadań. Jednak w sytuacji korelowania większej liczby równorzędnych profili, ze względu na wykładniczo rosnącą z liczbą profili liczbę możliwych korelacji, zadanie staje się trudne. Zaproponowane rozwiązanie ograniczenia czasu poszukiwania najlepszej korelacji wykorzystuje metodę Monte Carlo. Otrzymany wynik korelowania, aczkolwiek niekoniecznie najlepszy, najprawdopodobniej będzie bardzo bliski optymalnej korelacji. Jakość korelacji mierzona jest za pomocą współczynnika niepodobieństwa próbek. Końcowy wynik działania omawianego programu przedstawiany jest w postaci

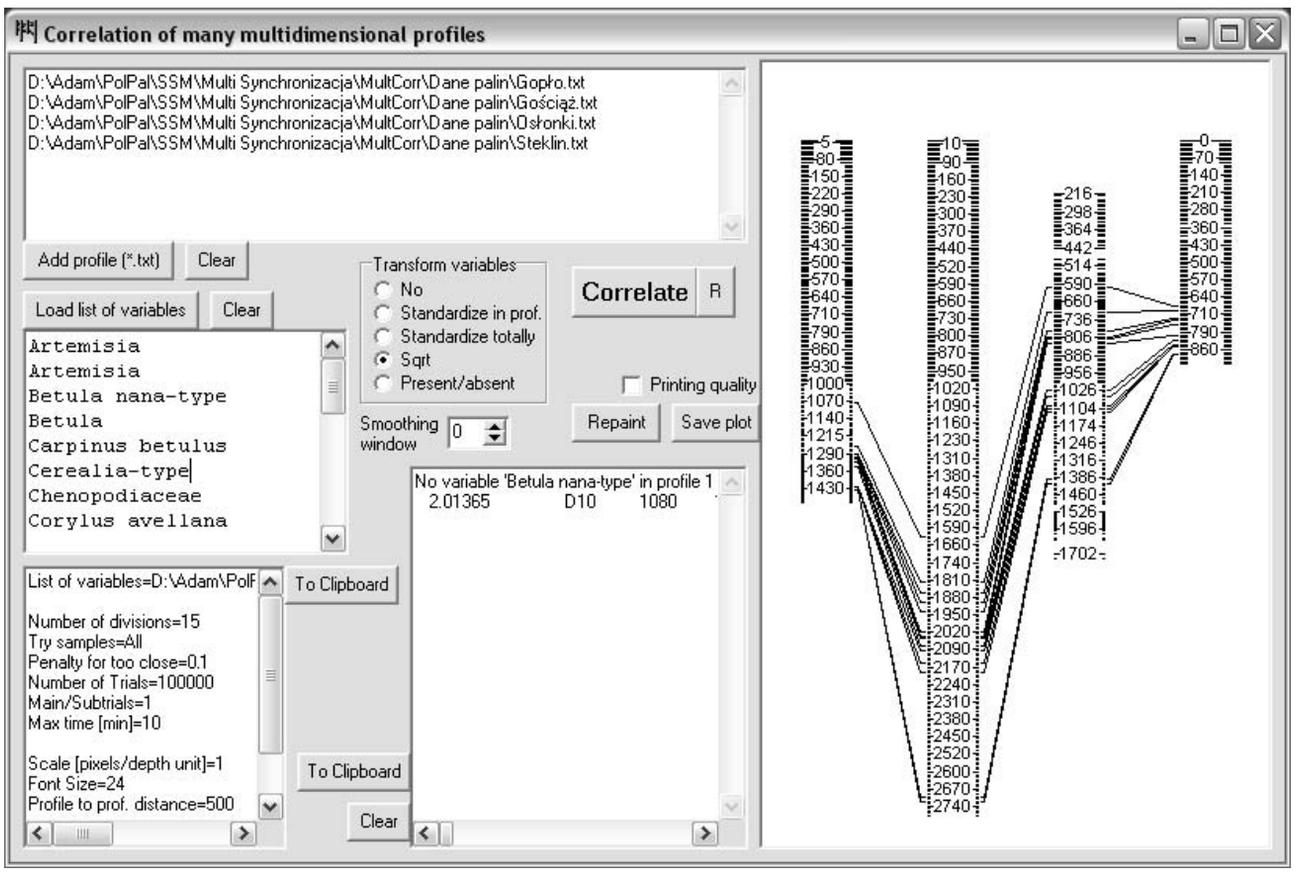

Fig. 13. The layout of the MultCorr application

graficznej, w postaci pewnej (zadanej) liczby linii łączących podobne poziomy. Liczba korelowanych profili, próbek i zmiennych zależy jedynie od wielkości pamięci komputera. Czas obliczeń zawsze można dowolnie ograniczyć, jednak warto wtedy sprawdzić stabilność uzyskanego wyniku.