HOW MANY VARIABLES ARE SUFFICIENT FOR THE DETERMINATION DISTURBANCE OF CHOLESTEROL METABOLISM?

Zdeněk Půlpán
University of Hradec Králové, Pedagogical Faculty, Department of Mathematics

***Abstract**: The paper describes the methodology of the so-called advanced data mining with help the programme parcel STATISTICA on the example of a "healthy-ill" ensemble.*

***Keywords:** data mining, answer tree methodology, statistical modelling and decision, data management*

1. Introduction

Software means are quite varied nowadays and they offer a number of levels of analysis of data ensembles consisting of various types of variables (both nominal and metric). Analyses are connected in a different way with statistical softwares (SPSS, NCSS, STATISTICA, etc.) and contain various kinds of procedures (Clementine, Data Miner, Neural Works, Answer Tree, Regression Tree, etc.). The present author will demonstrate one of the analyses elaborated with the use of the above-mentioned methodology. The results can be compared with its "classic" form published in [Půlpán 2003] and applied to the identical data.

Paper [Půlpán 2002, 2003, 2004] discussed the methodology of diagnosis determination on the basis of the construction of a multidimensional mathematical-statistical model containing four basic variables: *LTH* (lathosterol), *SIT* (sitosterol), *CAM* (camposterol), and *TCH* (total cholesterol). The diagnosis was formulated in the alternatives healthy-ill in connection with cholesterol metabolism. The decision-making was based on a basic sample of 101 subjects ("healthy" as regards with cholesterol metabolism) and samples of altogether 189 patients with various impairments of cholesterol metabolism. It has been shown that the data under study make it possible to establish diagnosis with the use of statistical methods with a degree of uncertainty not exceeding 30% of wrong diagnoses. In the present paper, an attempt will be made to establish the same diagnosis, but with the use of different means.

To obtain a set of measured values of the above-mentioned variables in healthy subjects is relatively expensive. The present author thus thinks that it is appropriate to present their more detailed processing, the results and possibilities of which can inspire further research.

2. Analysis of the set of the "healthy" subjects using the method of principal components

In a number of analyses we often examine a large number of variables, which, according to our assumptions, may be connected with the phenomenon under study. As we do not know the degree of action of the individual variables on the phenomenon studied, we attempt to introduce into the analysis as many variables as possible. However, it complicates the analysis and therefore we endeavour to find objective reasons for a selection of a smaller number of variables, which would be sufficient for the description of the phenomenon under study. Two multidimensional methods are available for this purpose: the method of principal components and factor analysis. Both methods search for a smaller number of new (unmeasurable, latent) variables, explaining variability and dependence of the original (measured) variables and their linear combinations. All original measurable variables enter the analysis as equal (though it need not be so from the standpoint of meaning). Their interrelations are explained by the action of mathematically defined directly unmeasurable (latent) variables, which in the analysis of principal components are called *components*, in factor analysis, *factors*. In the analysis of principal components, mathematical formalism is constructed in such a way so that the new variables (components) may explain the *variability* of the original variables as much as possible; in factor analysis, so that the factors may reproduce the linear relationships of the original variables (their correlation matrix) as best as possible. It is advantageous to require the latent variables not to be able to correlate.

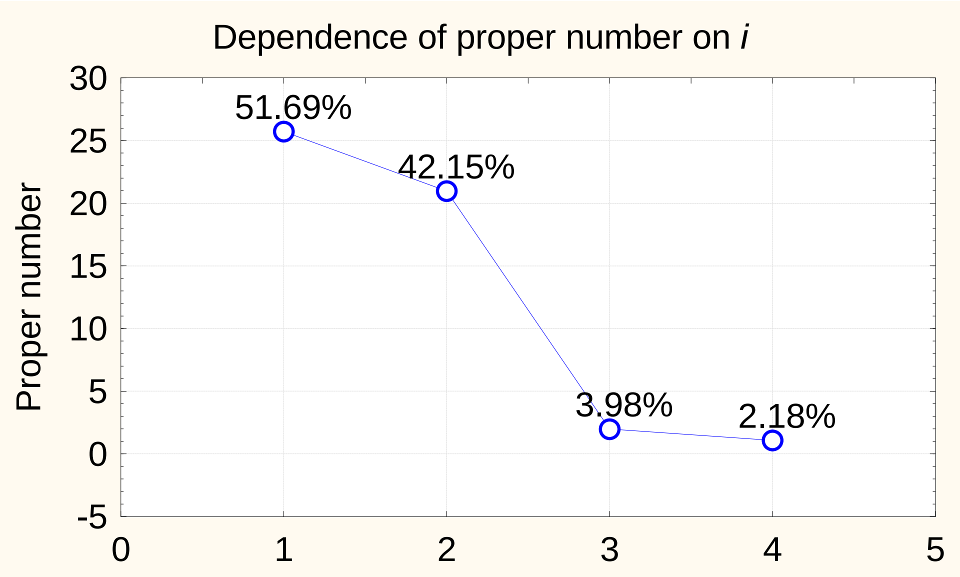
If we examine in *n* = 101 randomly selected "healthy" subjects the four signs *LTH*, *SIT*, *CAM*, and *TCH*, we have the measurement in the form of vectors ([1], Table 6)

$$\overrightarrow{x_i} = (LTH_i, SIT_i, CAM_i, TCH_i)', \quad i = 1, 2, \ldots, 101. \qquad (1)$$

For the above-mentioned vectors the selective covariantional matrix **C** and the characteristic numbers for it are estimated:

$$\mathbf{C} = \begin{pmatrix} 23.97 & 0.50 & 2.30 & 0.35 \\ & 6.52 & 8.60 & 0{,}68 \\ & & 18.06 & 0.79 \\ & & & 1.20 \end{pmatrix} \qquad (2)$$

$\lambda_1 = 25.72$, $\lambda_2 = 20.97$, $\lambda_3 = 1.98$, $\lambda_4 = 1.08$, explaining gradually 51.7 %, 42.2 %, 4.0 %, and 2.2 % from the total dispersion. This leads to the determination of *two* components, which represent the quality of substitution of measurable variables with the latent ones by about 94 %. It is graphically expressed in two-dimensional Graph 1, where to the order *i* of the characteristic numbers their value λ_i is assigned.



Graph 1: *Dependence of* l_i *on i* ($l_1 > l_2 > l_3 > l_4$)

Table 1 lists the component coordinates of the variables *LTH*, *SIT*, *CAM*, and *TCH* for two most important components. They are the coordinates of the characteristic vectors pertaining to the characteristic numbers λ_1, λ_2. (Graph 2 represents the variables under study in component coordinates.) By means of the data from Table 1, we can determine from the tetrad of the values of the coordinates $\overrightarrow{x_i}$ the couple of the main components (K_i^1, K_i^2) for *i* = 1,2…, 101 from the linear relations

$$K_i^1 = 0.80 \,.\, LTH_i + 0.26 \,.\, SIT_i + 0.54 \,.\, CAM_i + 0.04 \,.\, TCH_i - 13.25 \qquad (3)$$
$$K_i^2 = 0.60 \,.\, LTH_i - 0.39 \,.\, SIT_i - 0.69 \,.\, CAM_i - 0.03 \,.\, TCH_i + 4.50.$$

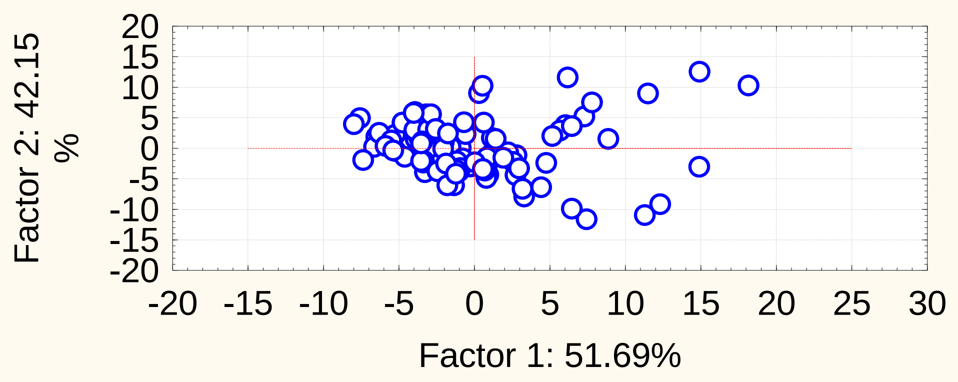
Table 2 assigns the component coordinates (K^1, K^2) to the individual cases of the "healthy" subjects according to the previous relations (3).

Tab. 1: Table of component coordinates.

Variable	Component	
	1.	2.
LTH	4.05 (0.80)	2.76 (0.60)
SIT	1.34 (0.26)	– 1.80 (–0.39)
CAM	2.74 (0.54)	– 3,18 (–0.69)
TCH	0.18 (0.04)	– 0.14 (–0.03)

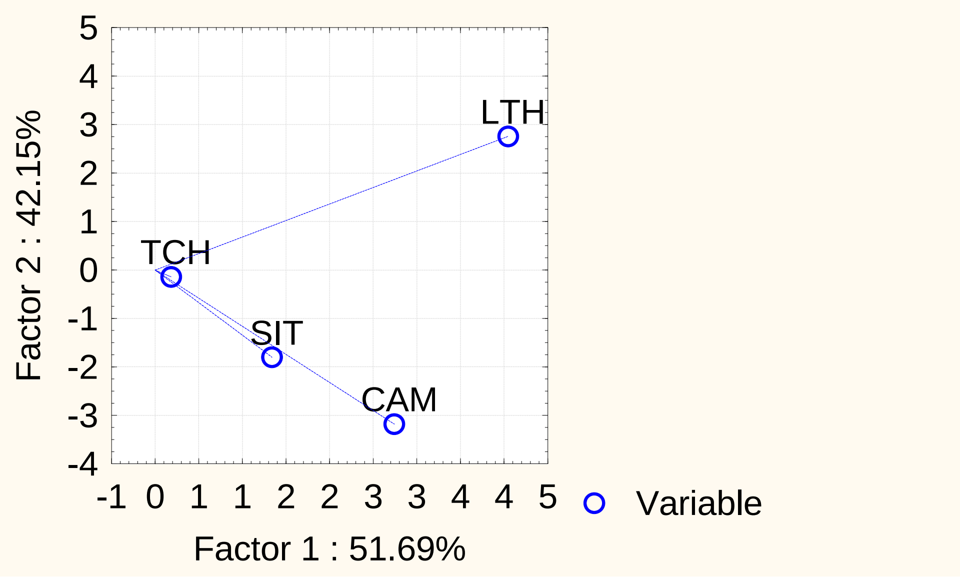
(They are the coordinates of the characteristic vectors of the first two characteristic numbers λ_1, λ_2, the normalized values are bracketed.)

The importance of the previous linear transformation of measured variables into the component ones consists in the fact that all measurements can be represented only in a two-dimensional graph with the axes of the coordinates K^1, K^2. In this new system of coordinates, all healthy subjects should cover a certain limited region, most probably a single cluster. The points lying rather far from the centre of gravity of the cluster correspond to the so-called remote values of measurement. The respondents corresponding to the remote values should be subjected to special examination (regarding the correctness of their inclusion into the "healthy" subjects). An idea about the arrangement of points (K_i^1, K_i^2), i = 1, 2,…, 101 in the ensemble of the healthy subjects under study can be obtained from two-dimensional Graph 2. In this Graph the measurements of respondents 83, 85, 3, 80, 19, … can be considered to be remote measurements.



Graph 2: The results of the measurements for the "healthy" subjects in two component coordinates

The data from Table 1 can be also transformed into two-dimensional Graph 3. For each of the variables *LTH*, *SIT*, *CAM*, and *TCH* there is a couple of coordinates which correspond to the first and second principal component. The representation of these variables in the coordinates formed by the values of the first and second component also gives an idea of the relation of the original variables in the new, component ones.



Graph 3: Projection of variables into the factor plane

If we use for component analysis the selective correlation matrix **R**

$$\mathbf{R} = \begin{pmatrix} 1.00 & 0.04 & 0.11 & 0.07 \\ & 1.00 & 0.79 & 0.24 \\ & & 1.00 & 0.17 \\ & & & 1.00 \end{pmatrix} \qquad (4)$$

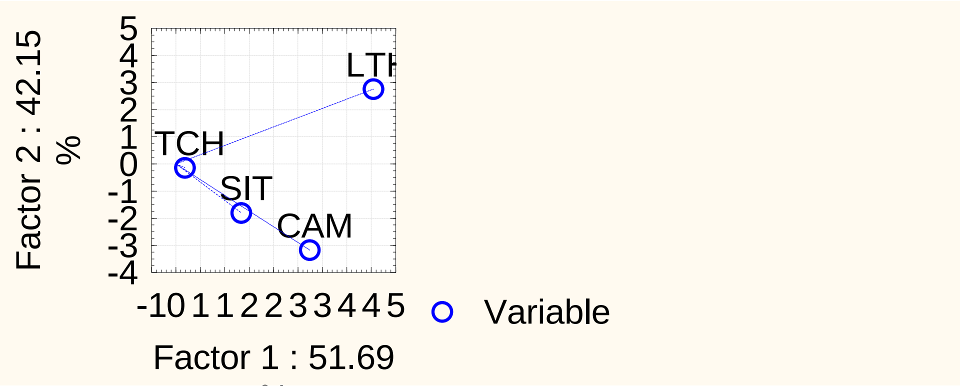
(the underlined selective correlation coefficients are statistically significant at a 5 % level of significance for the zero hypothesis of non-correlativeness), the characteristic numbers λ_1 = 1.90, λ_2 = 0.99, λ_3 =0.90, λ_4 = 0.20 are obtained, which gradually correspond to 47.6 %, 24.9 %, 22.5 %, and 5.0 of total dispersion. This corresponds to the quality of substitution of measurable variables with the latent ones by about 95 % in *three* component variables. Table 5 lists the corresponding component coordinates of the variables *LTH*, *SIT*, *CAM*, and *TCH*.

Tab. 2: Table of component coordinates under the assumption that they are based on the selective correlation matrix.

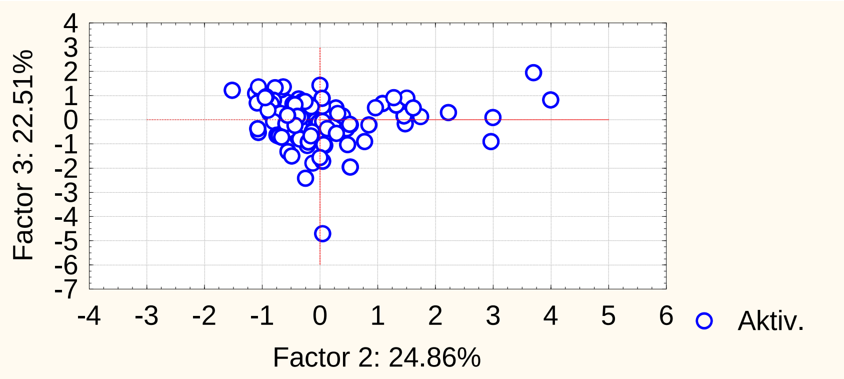
Variable	Component 1.	 2.	 3.
LTH	– 0.18	0.95	0.26
SIT	– 0.92	– 1.80	0.12
CAM	– 0.91	– 0.11	0.24
TCH	– 0.43	0.22	–0.87

As we can see from a comparison of Tables 1 and 2, the analysis of principal components based on the (selective) covariantional matrix **C** differs from the analysis of principal components based on the matrix of mutual (selective) correlations **R**. The correlation matrix is used as the introductory one in the case that the variables under study are of different nature and expressed in different units. In the present case the variables under study are expressed in identical units, but they measure different phenomena. Nevertheless, we attach greater significance to the component analysis from covariances.

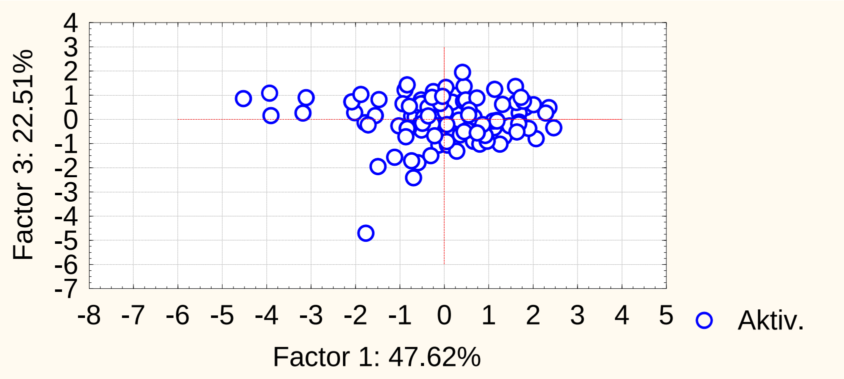
To examine the cluster of the "healthy" subjects in three component coordinates, the situations in three Graphs 4, 5, and 6 must be compared. The Graphs can be imagined as orthogonal projections of a three-dimensional cluster into some of the levels of coordinates.



Graph 4: Projection of a three-dimensional cluster into the level of the first and second factors (according to correlations)



Graph 5: Projection of a three-dimensional cluster into the level of the second and third factors (according to correlations).



Graph 6: Projection of a three-dimensional cluster into the level of the first and third factors (according to correlations).

The source data of the "healthy" subjects were processed also by the means of factor analysis (the method of the principal components without rotation). Two significant factors were extracted corresponding to two characteristic numbers $\lambda_1 = 1.905$ and $\lambda_2 = 0.994$. Their contribution to the result is evident in Table 3.

Tab. 3: Significance of characteristic numbers for factor analysis.

Characteristic number	% of total dispersion	Cumulative characteristic number	Cumulative %
1.905	47.62	1.905	47.6
0.994	24.86	2.900	72.5

Table 4 lists factor load, share of factors in communality, and coefficients of factor scores for individual variables. Table 5 then lists the residues of correlations which were not explained by the two above-mentioned factors.

Tab. 4: Basic results of factor analysis.

Variables	Factor loads		Communalities from the 1st factor	 from the 2nd factor	Coefficients of factor scores	
LTH	– 0.185	0.947	0.034	0.932	–0.097	0.953
SIM	– 0.923	– 0.180	0.852	0.884	– 0.485	– 0.181
CAM	– 0.912	– 0.116	0.832	0.846	– 0.479	– 0.116
TCH	– 0.432	0.226				

Tab. 5: Residues of correlations (underlined values represent significant differences between the actual correlations and their estimates with the use of factor analysis).

	LTH	*SIT*	*CAM*	*TCH*
LTH	0.07	0.04	0.05	– 0.23
SIT	0.04	0.12	– 0.07	– 0.12
CAM	0.05	– 0.07	0.15	– 0.20
TCH	– 0.23	– 0.12	– 0.20	0.76

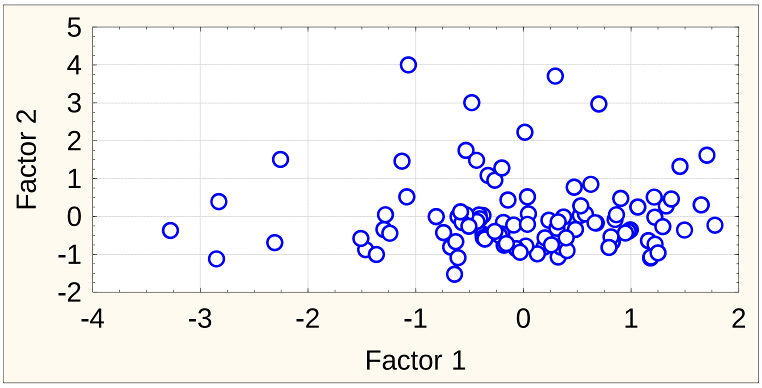
Similarly as in the analysis of principal components, also here it is possible to evaluate the remoteness of some measurements in the ensemble of the healthy subjects by means of the values of factor scores in two factors. The cluster of the “healthy” subjects is shown in Graph 13.

Table 4 can serve for the construction of the following scheme of the action of latent factors on manifest variables:

Factor 1 Factor 2

TCH

SIT, CAM *LTH*



Graph *7: Graph of the factor scores of the “healthy” subjects.*

3. The relation of the data of the ensemble of “healthy” subjects to “patients”.

Already at the beginning of the study we stressed that the aim of statistical survey in our case is to find an objective method for the classification of any individual into the ensemble of “healthy” subjects or “patients”. It cannot be clearly carried out categorically and therefore on the basis of the given data we attempt to determine the degree (perhaps as probability) of the classification of a subject into one of the two groups under consideration. The research method for the solution of this task is the answer tree methodology. On the basis of combinations of various statistical criteria, this procedure searches for the “optimal” classification into classes according to some categorial variable. It can be a new variable “condition”, which will have a value of 1 in the case of a healthy subject and 0 in the case of a metabolic disorder.

The result of the classification using the above-mentioned technique in an ensemble of 101 healthy subjects and 191 patients is shown in Graph 8. Graph 8 is a tree graph with three final nodes 3, 4, 5 (in the shape of a rectangle). Inside each rectangle in the left top corner there is the serial number of the node, in the right top corner there is the characteristic of the prevailing value of the pertinent “condition”, the degree of which is expressed by a histogram (dashed for “health”). The pertinent edges of the Graph are evaluated by the frequencies of the source elements of analysis. Between the two edges corresponding to the pertinent decision there is a brief statement about the condition of classification. On the basis of the conditions stated in the Graph, the “optimal” classification (from the standpoint of the programme *STATISTICA*) was performed and the results presented in Tables 6 and 7. The underlined values in Table 6 represent wrong diagnoses. We can see there 32.8 % and 21 %, respectively, of wrong predictions “healthy” or “ill”.

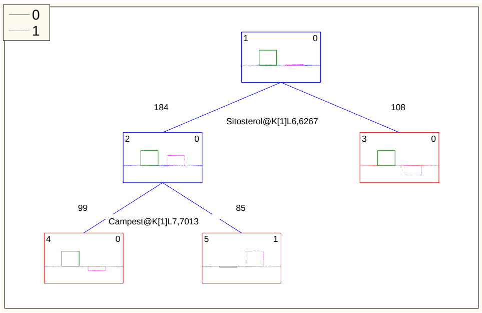
Tab. 6: *Result of discrimination analysis using the answer tree method.*

Prediction	Observation	
	0	1
0	164	43
1	27	58

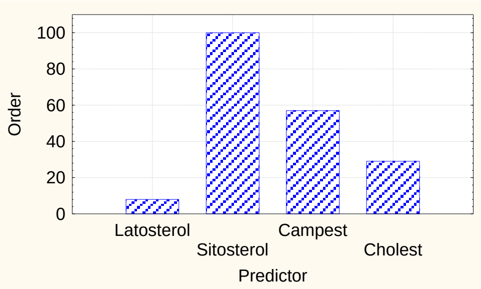
Tab. 7: Prediction with the use of an answer tree in contrast to “reality”.

Node	Left branch	Right ranch	Classes		Predicted class
			0	1	
1	2	3	191	101	0
2	4	5	100	84	0
3			91	17	0
4			73	26	0
5			27	58	1

The procedure makes it also possible to estimate the order of significance of predictors for the analysis. The variable *SIT* is of greatest significance (see Graph 9).

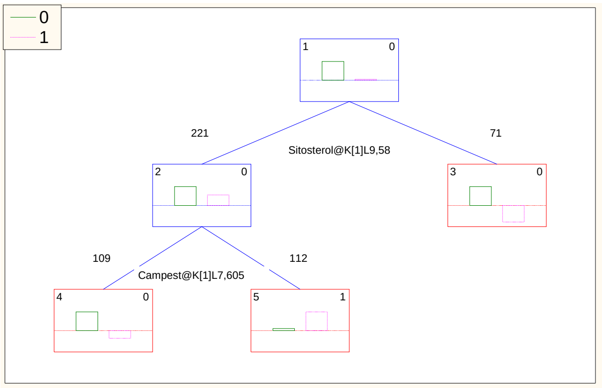


Graph 8: Answer tree for the classification into healthy – ill.



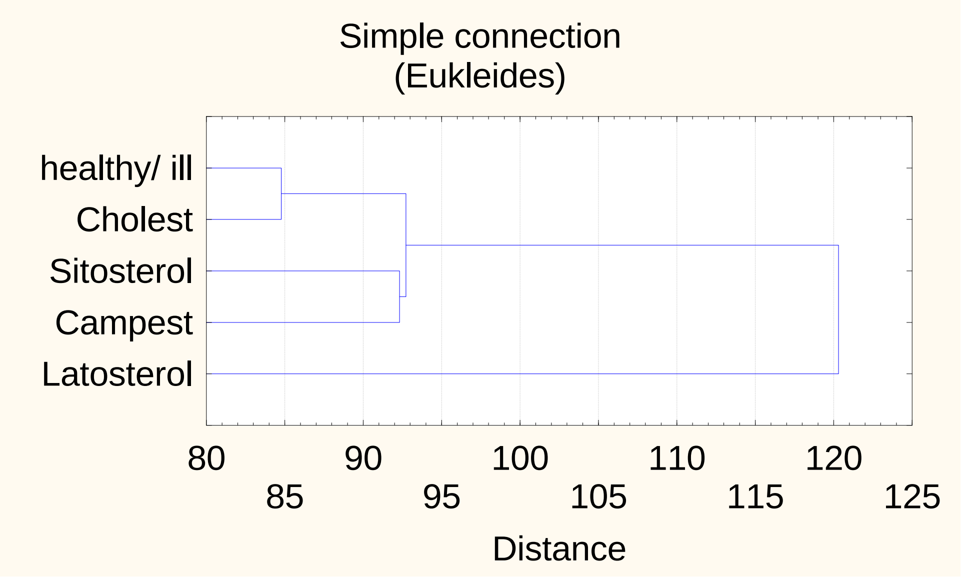
Graph 9: Order of significance of predictors for the analysis using an answer tree.

For the sake of information, it is necessary to state that answer trees may significantly differ when the method of the selection of branching or degree of agreement is changed (see Graph 10). That is why the method is called a "pilot" one.

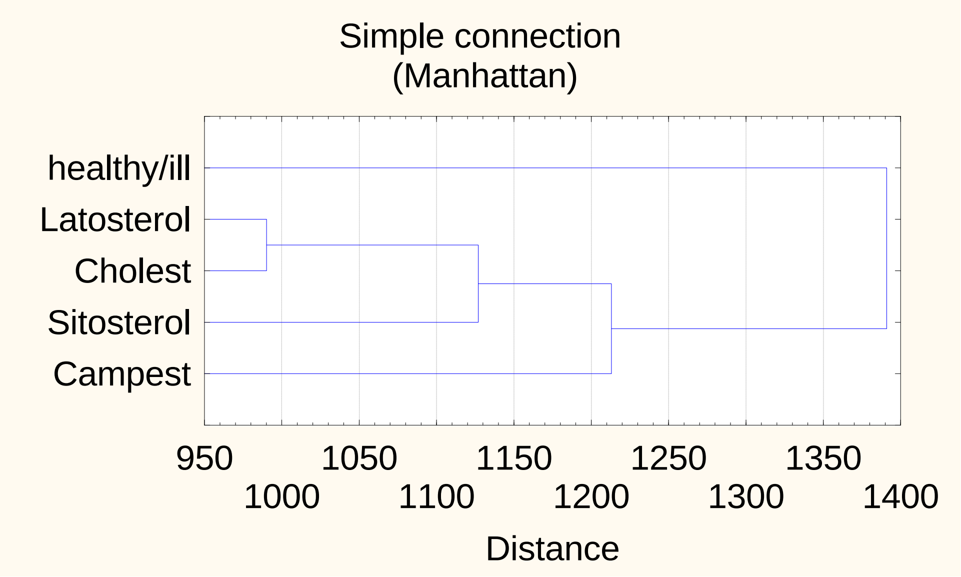


Graph 10: *An answer tree in the selection of branching using the method C&RT and the degree of agreement estimated by chi-quadrate.*

Also the so-called cluster analysis yields the result of the analysis in the form of a tree ([8] and [11]). It was employed also here. The ensembles of healthy subjects and patients were subject to hierarchical cluster analysis. The result unfortunately strongly depended on the selection of metrics (Graphs 11 and 12). Nevertheless, Graph 11 makes it possible to observe a closer connection of the cholesterol level and the diagnosis of a disease. But Graph 12 shows that classification into healthy subjects or patients is connected with the levels of all four sterols under study.



Graph 11: Hierarchical clustering in the Euclidean metric.



Graph 12: Hierarchical clustering in the Manhattan metric.

Now let us consider in the examined independent variables *LTH*, *SIT*, *CAM*, and *TCH* and the indicator of health *Y* the validity of the generalized linear model

$$g\,(E(Y/x)) = x' \,.\; b\;, \qquad (5)$$

where *g* is some function of the conditioned mean value of the random variable *Y*. In our case the vector of regressors $x' = (1, LTH, SIT, CAM, TCH)$. We search for a suitable vector of regression coefficients $b' = (b_0, b_1, \dots , b_4)$. We assume that n. v. *Y* possesses alternative division (in the respondent is healthy, then $Y = 1$, if he or she is ill, $Y = 0$). For alternative division $EY = p$ and if we select the function $g = g\,(p)$ in the form

$$g(p) = ln\ \frac{p}{1-p}\ = logit\,(p) \qquad (6)$$

Our model (18) is logistically regressive. Its parameters $b_0, b_1, \dots , b_4$ can be estimated by the method of maximal reliability, e.g., using the programme *STATISTICA.*

From the data of an ensemble of 292 “healthy” subjects and “patients”, the right side of expression (5) was obtained in the form

$$A = 1.174 + 0.012\ .\ LTH - 0.354\ .\ SIT + 0.221\ .\ CAM - 0.371\ .\ TCH \qquad (7)$$

The estimates of the coefficients b_0, b_1, … , b_4 from (7) possess 95 % intervals of reliability:

$$\begin{array}{lll} b_0 : (-2.39; 0.04) & b_1 : (-0.05; 0.03) & \\ b_2 : (0.24; 0.46) & b_3 : (-0.30; -0.14) & (8) \\ b_4 : (0.12; 0.62) . & & \end{array}$$

According to (5) and (6) we estimate

$$\ln \frac{p}{1-p} \sim A$$

$$p \sim \frac{e^A}{1+e^A} . \qquad (9)$$

It follows from estimates (8) that little influence of the variable *LTH* on n. v. *Y*. is possible.

The datum *p* can be then interpreted as the degree of “disease” of cholesterol metabolism. On the basis of *p* value each subject can be classified into some of the two groups, “healthy” or “ill”, by expert estimate, e.g., that in the case when $p > k_0{}^{kr}$, the subject is declared “healthy”, in the contrary case, “ill”. The value $p \rangle k_0^{kr}$, is considered with the use of a different technique (e.g., independent medical diagnosis).

Example: Let us have a subject *A*1 from the ensemble of “patients” with data (see [1]) *LTH* = 1.21; *SIT* = 6.80; *CAM* = 4.62; *TCH* = 4.98. According to (5), (6) and (7) then we have

$$A =. 1.174 + 0.012 . 1.21 – 0.354 . 6.80 + 0.221 . 4.62 = 0.371 . 4.98 =. –2.045$$

$$p \approx \frac{e^A}{1+e^A} = 0.115.$$

For the first subject from the ensemble of the “healthy” subjects with the data

LTH = 5.36; *SIT* = 6.25; *CAM* = 9.38; *TCH* = 3.59

we have A =. –0.233, $p \approx 0.442$.

Subject *A*1 is evaluated by logistic regression as “healthy” with the probability 0.115 and the first subject from the group of “healthy” subjects as “healthy” with the probability 0.442. ■

Table 8 lists the evaluation of diagnoses according to a priori values *Y* and the values obtained by logistic regression (of a posteriori values *Y*).

Tab. 8: Successfulness of diagnosis using logistic regression (k_0^{kr} = 0.5).

	A posteriori *Y*		
A priori	0	1	% correct
0	157	34	82.2
Y			
1	57	44	43.6

Table 8 shows that logistic regression estimates “disease” much better in the case of suspected disease than “health” in those who consider themselves to be healthy. (Is it the case also with the diagnoses realized outside logistic regression?)

We have also employed neural networks to analyze the sample of “healthy” subjects and “patients”. We wanted to find whether from the data of the variables *LTH*, *SIT*, *CAM*, and *TCH* it is possible to optimally divide by the algorithm of some of possible neural networks a sample of subjects into two groups (clusters), which could be interpreted in such a way that one of the groups would consist prevalently of healthy subjects, the other of patients. The procedure Intelligent Problem Solver was employed with the following types of networks.

1. MLP 4: 4 – 10 – 1 : 1
2. Linear 3: 3 – 1 : 1
3. Linear 2: 2 - 1 : 1
4. RBF 4: 4 – 10 – 1 : 1
5. RBF 4: 4 – 20 – 1 : 1.

The successfulness of these analyses can be judged from Tables 9a, b, c, d, e. The analysis seems to be very good for the given sample. The ability of the above-mentioned neural networks to correctly analyze the measurement of a subject who would not belong to the employed training ensemble, however, remains questionable.

Tab.9a): Successfulness of analysis using the network MLP (1).

A priori inclusion	A posteriori inclusion		Successfulness of inclusion
	0	1	successfulness %
0	143	48	74.9
1	20	81	80.2
Total	163	129	292

Tab. 9b): Successfulness of analysis using the network Linear (2).

A priori inclusion	A posteriori inclusion		Successfulness of inclusion
	0	1	successfulness %
0	139	52	72.7
1	23	78	77.2
Total	162	130	292

Tab. 9c): Successfulness of analysis using the network Linear (3).

A priori inclusion	A posteriori inclusion		Successfulness of inclusion
	0	1	successfulness %
0	140	51	73.3
1	24	77	76.2
Total	164	128	292

Tab. 9d): Successfulness of analysis using the network RBF (4).

A priori inclusion	A posteriori inclusion		Successfulness of inclusion
			successfulness %
0	0	1	77.0
1	17	84	83.2
Total	164	128	292

Tab. 9e): Successfulness of analysis using the network RBF (5).

A priori inclusion	A posteriori inclusion		Successfulness of inclusion
			successfulness %
0	152	39	79.6
1	15	86	85.1
Total	167	125	292

4. Conclusion

The programme parcel *STATISTICA* was used with advantage for the analysis. The analysis of relatively large ensembles of "healthy" subjects (altogether 101) and patients (altogether 191) could not bring a completely innovative methodology and the results of the performing of the concrete task of determination of cholesterol metabolism diagnosis also due to the fact that the statistical sample was not randomly selected from a larger and more homogeneous population and the number of analyzed variables was a priori limited to the data available to the author. In addition, such examination would require special preparation of the sample for analysis (particularly for the algorithms of neural networks). Nevertheless, we have outlined the possibilities and limits of the analyses used. Some mathematical methods would deserve to be repeated on a representative sample in cooperation with physicians. The author is a mathematician, not educated in medicine, and therefore he could bring some more courageous procedures into analyses. The kind reader will certainly compare the solution of the task with the "classic" technique. In addition, the methodology of data mining is characterized by effort to represent the results in tables and graphs in relatively hidden formal-mathematical algorithms and thus making it possible for non-mathematicians to use complex techniques.

References:

[1] PŮLPÁN, Z.: K formální definici nemoci. Acta medica (Hradec Králové) SUPPL 2003; 46 (1 – 2): 79 – 99

[2] HOSMER, D., W.: Applited Logistic Regression, Sekond Ed., J. Wiley, Canada, 2000

[3] SPANOS, A.: Probability Tudory and Statistical Inference, Cambridge Univ. Press 1999

[4] HÖPPNER, F., KLAWONN, F., KRUSE, R., RUNKLER, T.: Fuzzy Cluster Analysis, J. Wiley, England 2000

[5] ISTAC, J.: Mathematical Modeling for the Life Sciences, Springer, Berlin – Heidelberg, New York 2005

[6] AGRESI, A.: Categorical Data Analysis, Wiley, New York 1990

[7] HEBÁK, P. A KOL.: Vícerozměrné statistické metody, Informatorium, Praha 2005

[8] PŮLPÁN, Z.: Shluková analýza a její aplikace, Acta medica (Hradec Králové) , Suppl. 2002, 45 (1), 25 – 43

[9] KLASCHKA, J., ANTOCH, J.: Jak rychle pěstovat stromy, ROBUST´96, Sborník letních a zimních škol JČ(S) MF 2001

[10] ANTOCH, J.: Klasifikační a regresní stromy, ROBUST´88, Sborník letních a zimních škol JČ(S)MF 2001

[11] PŮLPÁN, Z.: K problematice zpracování empirických šetření v humanitních vědách, Academia, Praha 2004

[12] PŮLPÁN, Z.: Informoj por svaga aro en starigo de diagnozoj de malsanoj (Informations for fuzzy sets in illness diagnostics), grkg/ Humankybernetik, Band 41, Heft 4, 2000, S.167-177

[13] PULPÁN, Z.: Utilization of certain method from the field of data mining, v tisku

Contact address:

prof. RNDr. PhDr. Zdeněk Půlpán, CSc.
Hradec Králové, Pedagogická fakulta, Katedra matematiky
Rokitanského 62
500 03 Hradec Králové 3
Email: zdenek.pulpan@uhk.cz