PRACE NAUKOWE AKADEMII EKONOMICZNEJ WE WROCŁAWIU
Nr 1162 — 2007
Applications of Mathematics and Statistics in Economics

Petr Vlach, Miroslav Plašil

University of Economics, Prague, Czech Republic

VISUALIZATION OF MULTIVARIATE DATA

1. Introduction

Using multivariate data becomes very frequent in different fields of the economic research. We can meet them not only in the usual social or economic research, but also in marketing or sales studies, in the public sector research, etc. This general tendency is clear – describe the objects of interest as precisely as possible. The huge mass of new information brings together an increasing interest of researchers to cover all the complexity of the data. The information is often hidden in the data and not visible at the first sight.

The goal of data visualization is to provide a researcher with an insight into the complex phenomena of interest. Visualization techniques make use of the fact that human is very strong at the visual cognition, that enables him to grasp intuitively the interactions between variables (i.e. underlying structure of the data). Results of visualizations are very well suited to deliver useful information even to non-specialists.

However, the data obtained in many fields of actual research frequently pertain to a highly multivariate nature. With more than 3 variables one's ability to understand the interactions becomes severely limited by human's innate spatial perception and their visual *gift* to see data patterns inexorably fades away.

Investigating multivariate data in just one or two dimensions, as standard visualization tools (such as scatter plots and alikes) suggest, can be fairly misleading. Therefore we try to introduce some newer techniques for visualizing aspects of higher dimensional data sets: Bertin permutation matrices, parallel coordinates and RADVIZ.

By definition, all these techniques are exploratory, as they try maximally enhace one's understanding of the information in the data set. Multivariate visualization techniques summarize large amount of data into smaller and more clear repre-

sentations. They do not calibrate models, nor they test any hypothesis. Instead, they provide simple tools to interrogate the data in question. Even though different in nature, examined visualization techniques bear some interesting resemblance to the classical multivariate methods.

Please note that visualization differs substantialy from presentation graphics: the former is used to find meaningful patterns in unexplored data, whereas the latter adresses the information that is already understood.

2. Bertin (permutation) matrix

The Bertin matrix is a graphic method to investigate patterns in a set of quantitative data. In a loose sense, the Bertin matrix is a matrix of displays, where real values of an initial data matrix are transformed into graphical symbols of own choice (usually bars).

To make things more concrete, for each variable draw a bar chart of variable value by case and highlight all bars exceeding some threshold. As the following step, arrange the bar charts as a row of a matrix. This transformation of numerical values into a matrix of simple graphic elements makes the structure of a data set immediately visible.

It is obvious that rearrangements of a matrix such as row and column permutations do not mean any information lose, yet they enable to clean-up the initial matrix to get some more informative pattern. Even using *ad hoc* permutations one can identify patches that are ready for interpretation. Of course, with larger data sets *ad hoc* permutations become rather ineffective.

To formalize Bertin matrices and to sketch Bertin's strategy let us start with data matrix $\mathbf{X}_0$ (m variables and n cases). The matrix $\mathbf{X}_0$ is used to get a matrix of displays. For simplicity reasons we can identify $\mathbf{X}_0$ with the initial graph. To rearrange the initial graph we have a group of permutations Π (product of row and column permutations). We use $\mathbf{X}$ to denote a matrix after permutations, i.e. $\mathbf{X} = \pi\mathbf{X}_0$ for some permutation π and $\mathbf{X}[i,j] = \pi\mathbf{X}_0[i,j]$. To increase semantic contrast of visualization, it is useful to highlight information defined by $\mathbf{X}_0(i,j) > \text{mean}(\mathbf{X}_0\{.,j\})$.

Of course, not every permutation would lead to some understandable pattern. Searching for interesting patterns in data should reflect our explorative strategy. In a bid to search for data patterns in an effective way we introduce a *purity function* $\Phi = \Phi(\mathbf{X})$ that measures simplicity of the Bertin plot. Formally we look for optimal arrangements, that is a group of permutations π^* maximizing $\Phi(\pi\mathbf{X}_0)$. It implies that to find some informative pattern is a kind of optimization problem. Its computational intensiveness depends crucially on the complexity of a purity function. In real applications we are not limited to only one purity function, however the most common functions belong to the *arranging* purity function family: given row and column distances, we define the purity as the sum of distances of adjacent

rows/columns. In practice we usually restrict ourselves to separate optimization arrangement problems for rows and columns[1].

The follow-up statistical analysis whether π^* reflects any true information or it is just an expression of random effects may be based on the permutation theory.

3. Parallel coordinates

Parallel coordinates, pioneered by A. Inselberg in 80's, belong arguably to the most intuitive visualization techniques. In this visualization, n-dimensional vector is projected onto a plannar diagram. Unlike to the traditional Cartesian coordinate system, final display is obtained by taking the dimensions as vertical axes arranged parallelly to each other. Thus, if the original vector is written as $\mathbf{x} = (x_1, x_2, ..., x_n)$ then, its parallel coordinate representation is a polyline connecting the points $\{1, x_1\}$, $\{2, x_2\}$, ..., $\{n, x_n\}$. Each polyline represents one point of a multidimensional space and could be thought as a *profile* of a given case. The actual shape of a profile bears usually enough information to draw conclusions about structure of the data set.

Parallel coordinates serve as a useful tool to detect outliers (through a *suspicious* or *unique* shape of the profile), to identify relations between variables and/or to identify interpretable clusters (e.g. set of lines with a similar gradient indicate that data correlate positively). However, the use of parallel coordinates is still opened to a broader range of data analysis (e.g. problems of discrimination between two or more groups).

As the size of the data set increases, the structure in data may become obscure for the visual clutter, which makes it hard for the viewer to discern patterns and relationships. Therefore, this visualization technique gains its real strength if used in an interactive environment (which is typical for a modern multivariate visualization) when a researcher can vary the dimensions order and work with selected subsets of the data, usually on the basis of one particular variable.

4. RADVIZ (Radial Coordinate Visualization)

RADVIZ method maps a set of n-dimensional points onto two-dimensional space. To explain nature of the RADVIZ approach one usually resorts to its physical analogy. Suppose there are n points equally spaced around the circumference of the circle. These points are called dimensional achors and are denoted S_1 to S_n. Now suppose a set of n springs emanating from each of these points, and that all of the springs are attached to the other end to a puck.

Finally, assume the stiffness constant of the j-th string is x_{ij} for one of the data points i. If the puck is released, it would, by physical laws, head for equlibrium posi-

[1] For more detailed mathematical treatment see [10] and the references therein.

tion, the coordinates of this position $(u_i, v_i)^T$ say, are the projection of the n-dimensional point into two-dimensional space. To get a visualization of n-dimensional data set via RADVIZ, we compute and plot positions (u_i, v_i) for each case.

To put some additional light on the RADVIZ projection, consider the forces acting on the puck. For a given spring, the force acting on the puck is the product of the vector spring extension and the scalar stiffness constant. To determine resultant force and its direction we have to sum up all forces acting on the puck. It follows that in the point of equilibrium where there are no additional forces acting on the puck, this sum will be zero. Thus, denoting the position vectors of S_1 to S_n by S_1 to S_n and putting $u_i = (u_i, v_i)^T$ we get

$$\sum_{j=1,m} \left(S_j - u_i\right) x_{ij} = 0, \tag{1}$$

which may be solved for u_i by

$$u_i = \sum_{j=1,m} w_{ij} S_j, \tag{2}$$

where

$$w_{ij} = \left(\sum_{j=1,m} x_{ij} \right)^{-1} x_{ij}. \tag{3}$$

Now it is clear that for each case i, u_i is simply a weighted mean of the S_j's whose weights are the n variables for case i normalized to sum to one. Note that this normalization operation makes the projection nonlinear.

The nonlinear trasformation of the data preserves certain symmetries. Some features of this visualization include above all:

- points with approximately same dimensional values lie close to the center,
- points which have one or two coordinate values greater than the others lie closer to the dimensional anchors of those dimensions,
- the position depends heavily on the layout of the particular dimensions around the circumference of the circle.

These features, for example, imply that observations in which all variables have very high values will be mapped onto the same point as observations in which all variables take on a very low constant value. This rather strange and undesirable property of RADVIZ sometimes makes the visualization hard to interpret. Simulations with real data sets, however, suggest that RADVIZ rarely fails to provide reasonable plots of explored data structure even in the presence of such pathological cases.

5. Multidimensional Scaling

Multidimensional Scaling (MDS) is a set of mathematical techniques that enable a researcher to uncover the hidden structure of data by its geometrical representation. MDS can be used in wide range of situations such as exploratory data analysis, data reduction or visual clustering. The goal of the algorithm is to reproduce proximity matrix of multivariate data in a low dimensional space, as known as *configuration of points*.

In classical (or metric) multidimensional scaling the proximities are treated directly as distances. We start with an $n \times n$ distance matrix $\mathbf{D}$. The goal is to find n points in k dimensions such that the interpoint distances d_{ij} in the k dimensions are approximately equal to the values δ_{ij} in $\mathbf{D}$. The solution can be found in two basic steps:

1. First we construct the double centered proximity matrix $\mathbf{B}$:

$$\mathbf{B} = -\frac{1}{2}\left[\mathbf{I} - \frac{1}{n}\mathbf{i}\mathbf{i}^T\right]\mathbf{D}^2\left[\mathbf{I} - \frac{1}{n}\mathbf{i}\mathbf{i}^T\right], \tag{4}$$

where $\mathbf{I}$ is an identity matrix, and $\mathbf{i}$ an unity vector.

2. Since $\mathbf{B}$ is symmetric and usualy positive definite we can find the coordinates of points from the spectral decomposition of $\mathbf{B}$:

$$\mathbf{B} = \mathbf{V}\Lambda\mathbf{V}^T, \tag{5}$$

where Λ is the matrix of eigenvalues of $\mathbf{B}$ and $\mathbf{V}$ is the matrix of their corresponding eigenvectors.

From the equation above, we can write

$$\mathbf{B} = \mathbf{V}\Lambda^{1/2}\Lambda^{1/2}\mathbf{V}^T \tag{6}$$

and hence

$$\mathbf{B} = \mathbf{X}\mathbf{X}^T, \tag{7}$$

where

$$\mathbf{X} = \mathbf{V}\Lambda^{1/2}. \tag{8}$$

The columns of $n \times k$ matrix $\mathbf{X}$ are the coordinates of points in k-dimensional space, for which the interpoints distances d_{ij} corresponds to the distances δ_{ij}.

In nonmetric MDS we find a low-dimensional configuration of points such that the ranking of the distances corresponds to the ordering of the original dissimilarities. We usually start with some starting configuration of points and their interpoint distances. In case of lack of monotonicity (interpoint distances ranking does not match the ranking of the original distances) a suitable transformation is estimated by *mono-*

tonic regression. Then we estimate new values of d^*_{ij} (called disparities) under the assumption to minimize some of STRESS criterions (goodness-of-fit measures).

6. Simulation[2]

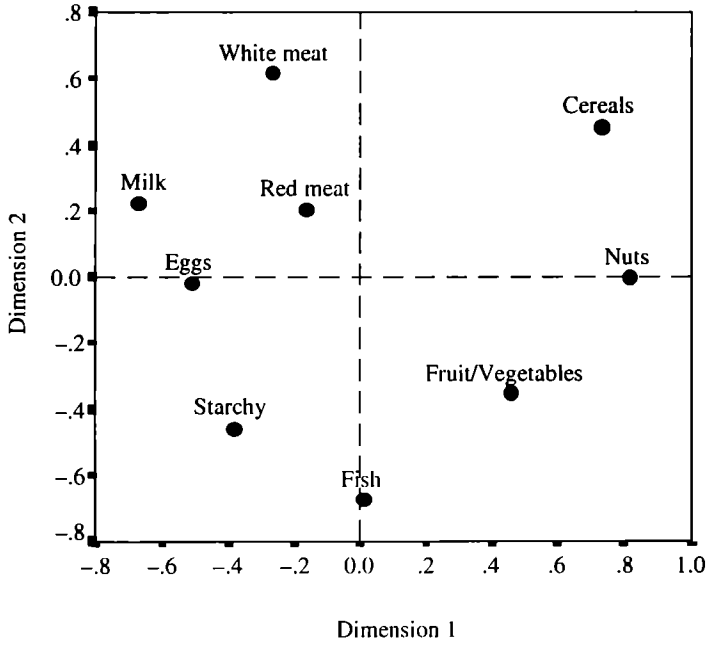
For the illustration of the algorithms described above we use the dataset taken from [3]. The data in Table 1 contain measurements of protein consumption in 25 European countries for 9 goods groups.

Table 1. Protein data

Country	Red meat	White meat	Eggs	Milk	Fish	Cereals	Starchy	Nuts	Fruit / vegetables
ALB	10.1	1.4	0.5	8.9	0.2	42.3	0.6	5.5	1.7
AUS	8.9	14	4.3	19.9	2.1	28	3.6	1.3	4.3
BEL	13.5	9.3	4.1	17.5	4.5	26.6	5.7	2.1	4
BUL	7.8	6	1.6	8.3	1.2	56.7	1.1	3.7	4.2
CZE	9.7	11.4	2.8	12.5	2	34.3	5	1.1	4
DEN	10.6	10.8	3.7	25	9.9	21.9	4.8	0.7	2.4
GER_east	8.4	11.6	3.7	11.1	5.4	24.6	6.5	0.8	3.6
GER_west	11.4	12.5	4.1	18.8	3.4	18.6	5.2	1.5	3.8
FIN	9.5	4.9	2.7	33.7	5.8	26.3	5.1	1	1.4
FRA	18	9.9	3.3	19.5	5.7	28.1	4.8	2.4	6.5
GRE	10.2	3	2.8	17.6	5.9	41.7	2.2	7.8	6.5
HUN	5.3	12.4	2.9	9.7	0.3	40.1	4	5.4	4.2
IRE	13.9	10	4.7	25.8	2.2	24	6.2	1.6	2.9
ITA	9	5.1	2.9	13.7	3.4	36.8	2.1	4.3	6.7
NET	9.5	13.6	3.6	23.4	2.5	22.4	4.2	1.8	3.7
NOR	9.4	4.7	2.7	23.3	9.7	23	4.6	1.6	2.7
POL	6.9	10.2	2.7	19.3	3	36.1	5.9	2	6.6
POR	6.2	3.7	1.1	4.9	14.2	27	5.9	4.7	7.9
ROM	6.2	6.3	1.5	11.1	1	49.6	3.1	5.3	2.8
RUS	9.3	4.6	2.1	16.6	3	43.6	6.4	3.4	2.9
SPA	7.1	3.4	3.1	8.6	7	29.2	5.7	5.9	7.2
SWE	9.9	7.8	3.5	24.7	7.5	19.5	3.7	1.4	2
SWI	13.1	10.1	3.1	23.8	2.3	25.6	2.8	2.4	4.9
UK	17.4	5.7	4.7	20.6	4.3	24.3	4.7	3.4	3.3
YUG	4.4	5	1.2	9.5	0.6	55.9	3	5.7	3.2

First we try multidimensional scaling to discover the main patterns in the data. We explore both variables and cases. The resulting configurations in the first and second dimensions are given in Fig. 1.

[2] Computed with SPSS 11.0, Visulab 4.0. and Qt Orange Canvas.



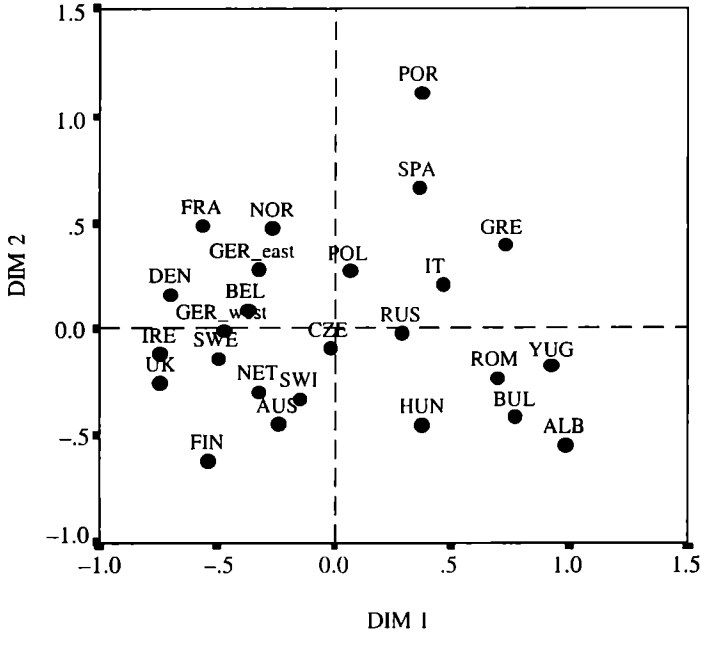
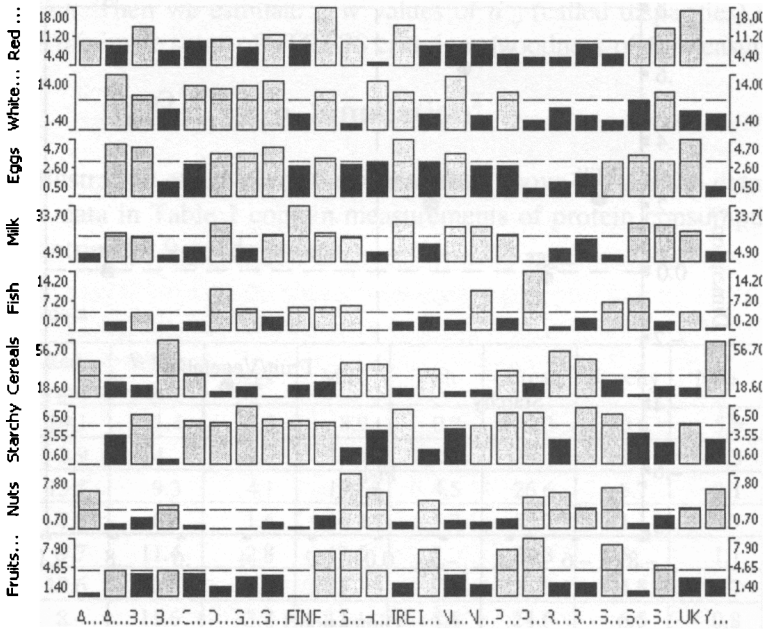


Fig. 1. Configuration of points of the protein data



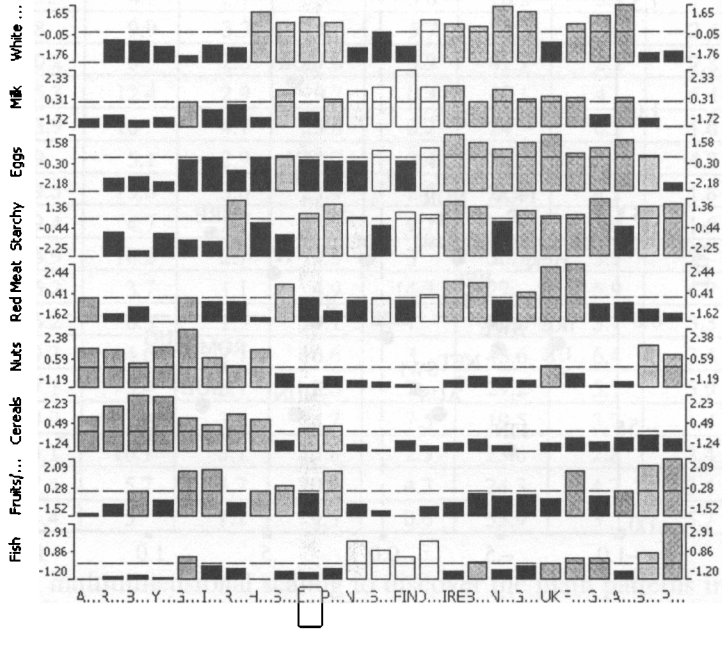


Fig. 2. Bertin permutation matrix of the protein data

From Fig. 1 we clearly see that the first component separates *Red* and *White meat*, *Eggs*, *Milk* and *Starchy* food from *Cereals*, *Nuts*, *Fish* and *Fruit/vegetables*. This separation is logic and can be explained as two main factors: livestock production and the plant production. From the second configuration we see the clustering of objects (countries) into five clusters: YUG-BUL-ROM-HUN-ALB, CZE-RUS-POL, ITA-SPA-GRE, FIN-AUS-SWI-NET-SWE-UK-IRE and FRA-NOR--GER-BEL-DEN.

Now let us have a look on the data using Bertin permutation matrix. The input set is in the shape of row data. After a few permutations we get result given in Fig. 2.

The permutations yield in more homogenous structure, from which we can conclude that there are three basic groups of protein consumption: *White meat*, *Red meat*, *Milk*, *Eggs* and *Starchy* followed by *Nuts* and *Cereals* in the second group and *Fish* and *Fruit/vegetables* in the last one. Shades group together the countries. As we can see, there is similar grouping as we saw previously in Fig. 1.

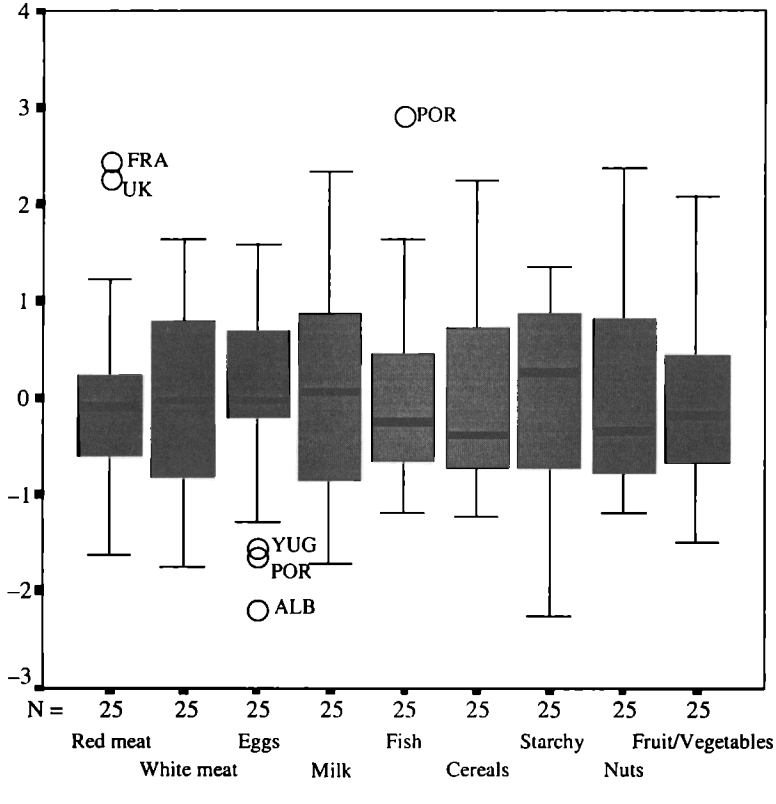


Fig. 3. Boxplot of the protein data

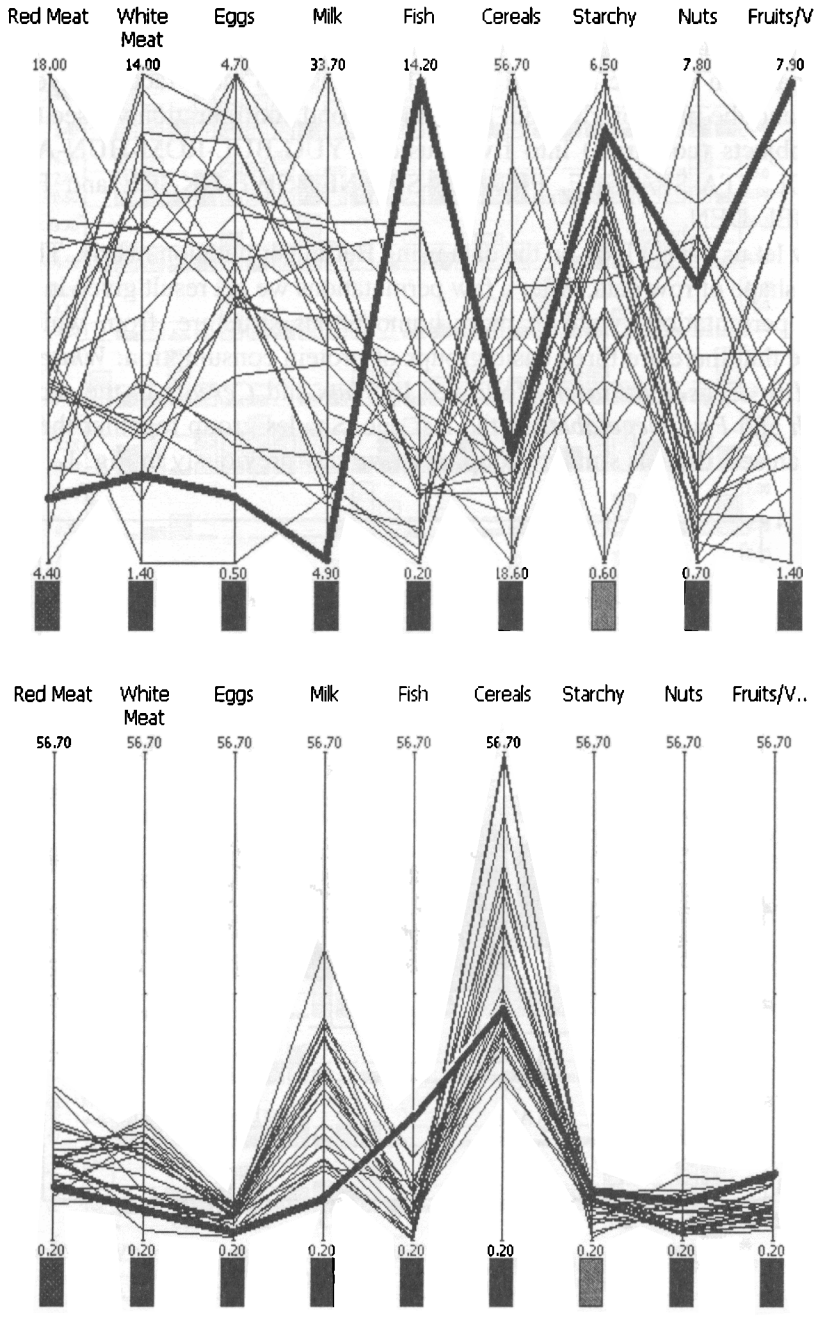


Fig. 4. Parallel coordinates for *Portugal* (bold) – original and rescaled values

The main reason of using Bertin permutation matrix is the possibility to get deeper in the data. Bertin matrix is effective not only for small data sets, but can be applied to larger samples (but not too large, because results could become fairly bustling). This algorithm is not giving information about a particular case, but gives the summary view on the data.

From a different point of view we try to apply the parallel coordinates on the same data set. The parallel coordinates algorithm (a.k.a. *parallel profiles*) is used to discover the general patterns in your data and is very suitable tool for discovering outliers or extremes. They appear as single curves sticking out of common behaviour of the others.

Let us have a look at usual way for identifying outliers *via* box plots. As we can see, there are three variables (standardized to z-scores) with outliers – *Red Meat*, *Eggs* and *Milk*. From Fig. 3 we simply identify which cases are affected. But we hardly see its values on the other variables. We only know if they are extremes or not.

To resolve this task, just use parallel coordinates and plot the graph, as depicted on Fig. 4. The numbers above and below the graph show the maximum and minimum values in the data. The values are rescaled to be mutually comparable.

Bold line shows *Portugal* (POR), formerly identified as the outlier on the variable *Fish*. As we can see, this country is high on *Fish*, but very low on *Milk* – not seen in Fig. 3.

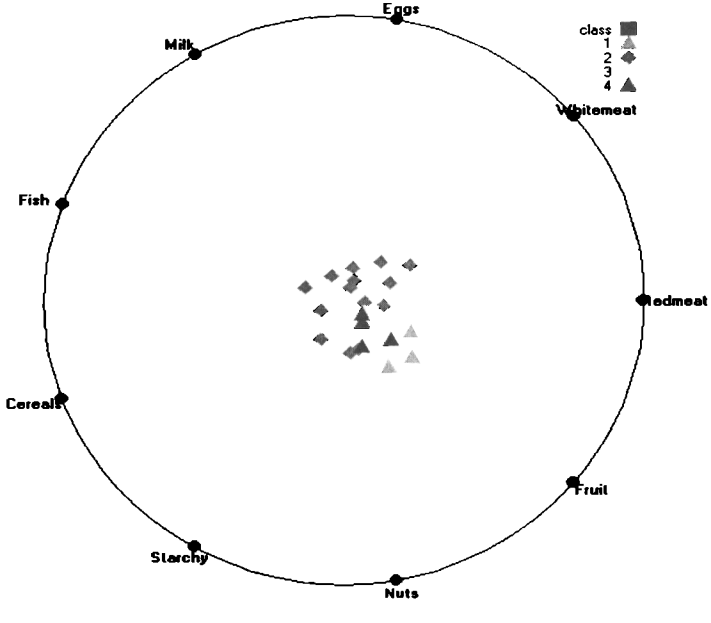


Fig. 5. RADVIZ of the protein data

Finally RADVIZ gives very similar results to the previous methods as we can see in Fig. 5. Unfortunately the graphics does not support labelling so the points were shaded subjectively according to the mapping to get interpretable results.

In comparison with other methods of exploration of multivariate data all the mentioned visualization methods provide an easy, quick and comparable relevant information. They can definitely help you in:

- segmentation,
- clustering,
- exploring data.

7. Conclusion

Modern visualization provides easy tools to solve classical multivariate problems. Without using complicated mathematical apparatus, they convey very similar information as more sophisticated traditional multivariate methods. They do not calibrate model but can be very instrumental in formulating the hypothesis for subsequent testing.

This contribution was supported by Internal Grant Agency (IGA) at the University of Economics, Prague, grant No. 410054.

References

[1] Cox M.J., Cox T.F., *Multidimensional Scaling*, Chapman & Hall, London 2001.

[2] Greenacre J., Blasius J., *Correspondence Analysis in the Social Sciences, Recent Developments and Applications*, Academic Press, London 1994.

[3] Rencher A., *Methods of Multivariate Analysis*, Wiley Interscience Publication, New York 2002.

[4] Hebák P., Hustopecký J., *Vícerozměrné metody s aplikacemi*, SNTL, Praha 1987.

[5] Brunsdon C., Fotheringham A.S., Charlton M.E., *An Investigation of Methods for Visualising Highly Multivariate Datasets*, http://www.agocg.ac.uk/reports/visual/casestud/brunsdon/brunsdon.pdf, 1997.

[6] Hoffman P.E., *Table Visualizations: A Formal Model and Its Applications*. A dissertation thesis University of Massachusetts Lowell, http://home.comcast.net/~peh2.hoffman/tablevizx.pdf, 1999.

[7] Schmid C., Hinterberger H., *Comparative Multivariate Visualization across Conceptually Different Graphic Displays*, Institute for Scientific Computing, ETH Zürich, 1994.

[8] Wei Peng, *Clutter-based Dimension Reordering in Multidimensional Data Vizualization*, A thesis Worchester Polytechnic Institute, http://www.wpi.edu/Pubs/ETD/Available/etd-01115-222940/unrestricted/wpeng.pdf, 2005.

[9] *ATKOsoft S.A. Manual: Survey on Visualization Methods and Software Tools*, 1997.

[10] Falguerolles A. de, Friedrich F., Sawitzki G., *A Tribute to J. Bertin's Graphical Data Analysis*, http://www.statlab.uni-heidelberg.de/reports/by.series/beitrag.34.pdf, 1996.

[11] Brown M.P., Grundy W.N., Lin D., Cristianini N., Sugnet C., Furey T.S., Ares M., Hausller D., "Knowledge-based Analysis of Microarray Gene Expression Data by Using Support Vector Machines", *Proceedings of the National Academy of Sciences* 2000, 1, 262–267.

WIZUALIZACJA DANYCH WIELOWYMIAROWYCH

Streszczenie

Dane ekonomiczne są często wielowymiarowe. Jest to konsekwencją tendencji do opisywania badanych obiektów tak precyzyjnie, jak to możliwe. Wiele metod statystycznych jest stosowanych do jedno- lub dwuwymiarowych danych, dlatego też starano się przedstawić pewne aspekty eksploracji i wizualizacji danych wysoce wielowymiarowych. Konsekwentnie położono nacisk na pewnych interesujących związkach z tradycyjnymi metodami analizy wielowymiarowej, które ze swej natury nie służą do wizualizacji.