A C T A U N I V E R S I T A T I S L O D Z I E N S I S
FOLIA OECONOMICA 90, 1989

*Zdzisław Hellwig**, *Edward Nowak***

AN UNSUFFICIENT INFORMATION PROBLEM IN TAXONOMIC MODELLING

1. Let X denote a data matrix:

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{nm} \end{bmatrix}$$

where:

n - number of objects,
m - number of attributes (variables).

The element x_{ik} $(i = 1, 2, \dots, n,\ k = 1, 2, \dots, m)$ of this matrix is called an elementary information. The rows of this matrix are called vector - objects, the columns are called vector - attributes.

If each element of data matrix is known, then this matrix is called a complete data matrix. A matrix X is incomplete, if some elements of this matrix are unknown. Then the lack of elementary information is called a gap in data matrix.

2. If a gap in data matrix occurs, then its size should be analysed. One should assume the sufficient tolerance level in lack of information, say p%. Two basic conditions should be fulfilled:

* Professor at the Academy of Economics Wrocław.
** Lecturer at the Academy of Economics, Wrocław.

- in each row of data matrix no more than p% of elements are unknown,

- in each column of data matrix no more than p% of elements are unknown.

Three tolerance levels are proposed:

1) p = 10%, so called rigid tolerance level,

2) p = 20%, so called average tolerance level,

3) p = 30%, so called mild tolerance level.

If more than p% of values are missing, then the corresponding rows or columns are removed from data matrix. For the other rows or columns missing values may be completed.

3. Data should be collected by proper institutions. These institutions would also complete these data. In Central Statistical Office the data bank for Poland should be established. This bank would be available for any institution. For international comparisons the data bank in Cracow Academy of Economics would be used.

4. Missing values in data matrix may be completed by means of two groups of methods:

- methods with external information,

- methods without external information.

5. The methods with external information consist in the estimation of missing values using available data. Only some methods are admissible. The most important admissible methods are:

- interpolation of trends,

- estimation of missing values by means of regression equations,

- taxonomic method.

The taxonomic method consists in determining for each object for which missing values occur, his nearest neighbours with respect to all attributes, for which the data are available. For each attribute two nearest neighbours should be determined: "right - hand side neighbour" and "left - hand side neighbour". The average of the values for these neighbours is the estimate for missing value.

6. For methods with external information, the data corresponding to other than considered object are used. So in this sense the data are "polluted". In methods without extraneous information modified taxonomic procedures are used. Using the data

available for the considered object they make the modelling possible.

Here, two methods are proposed:

- modified Euclidean metric method,
- correlation metric method.

7. The modified Euclidean metric method makes possible to compare multi-attribute objects (that is multidimensional observations) in different spaces: $\mathfrak{X}^m$, $\mathfrak{X}^s$ $(m > s)$. For each object such vector space is determined, that the data matrix reduced in this space is complete. The distance between two objects is defined as:

$$d_{ij} = \sqrt{\frac{m}{s}} \cdot \sqrt{\sum_{k=1}^{s} (x_{ik} - x_{jk})^2} ,$$

where:

s - number of attributes, for which the data corresponding to both the ith object and the jth object are available.

The value $\sqrt{\frac{m}{s}}$ is a correction coefficient.

The distances between objects, for which all data are available, are determined according to the formula:

$$d_{ij} = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2} .$$

8. The correlation metric method consists in determining a correlation matrix R, whose elements

$$r_{kl} = r(X_k, X_l)$$

are calculated by means of these components of object vectors. The correlation metric is defined as:

$$d_{kl} = 1 - |r_{kl}|,$$

or:

$$d_{kl} = 1 - r_{kl}^2.$$

9. An example. Ten countries were chosen as objects: 1) Bulgaria, 2) Czechoslovakia, 3) France, 4) Spain, 5) GDR, 6) Poland, 7) West Germany, 8) Hungary, 9) Great Britain, 10) Italy.

These objects are compared with respect to six attributes:

X_1 - wheat yield in kg per ha,

X_2 - barley yield in kg per ha,

X_3 - potato yield in kg per ha,

X_4 - beef meat production in kg per 1 ha of farmland,

X_5 - pork meat production in kg per 1 ha of farmland,

X_6 - milk production in kg per 1 ha of farmland.

Values of attributes (except for x_{21} i.e. wheat yield in Czechoslovakia) are given in Table 1.

The distances between Czechoslovakia and some other objects were calculated on the basis of the attributes X_2, X_3, X_4, X_5 and X_6. The values of distances (before modification) are given as a vector

$$d_2^o = [1{,}390 \quad 0{,}000 \quad 1{,}432 \quad 1{,}839 \quad 1{,}277 \quad 1{,}132 \quad 2{,}077 \quad 1{,}297 \quad 1{,}575 \quad 1{,}004].$$

Since $m = 6$ and $s = 5$, then the correction coefficient

$$\sqrt{\frac{m}{s}} = 1{,}095.$$

Multiplying d_2^o by this coefficient we get

$$d_2 = [1{,}522 \quad 0{,}000 \quad 1{,}568 \quad 2{,}014 \quad 1{,}399 \quad 1{,}239 \quad 2{,}274 \quad 1{,}420 \quad 1{,}725 \quad 1{,}099].$$

Table 1

Standardized values of attributes

Objects	X_1	X_2	X_3	X_4	X_5	X_6
1. Bulgaria	0,343	0,516	-1,333	-0,953	-0,652	-1,044
2. Czechoslovakia	.	0,195	-0,404	0,154	0,011	0,006
3. France	0,716	0,619	1,423	0,344	-0,676	0,447
4. Spain	-1,851	-2,385	-0,599	-1,224	-1,079	-1,273
5. GDR	0,343	0,596	-0,419	-0,099	1,313	0,861
6. Poland	-0,845	-0,745	-0,839	-0,388	-0,487	0,105
7. West Germany	0,967	1,170	1,138	2,448	2,152	2,335
8. Hungary	0,160	-0,344	-0,419	-1,037	0,686	-0,811
9. Great Britain	1,333	0,951	1,797	0,095	-0,841	0,059
10. Italy	-1,173	-0,573	-0,314	0,462	-0,431	-0,472

The complete distance matrix is:

$$D = \begin{bmatrix} 0 & 1,522 & 1,847 & 1,936 & 1,735 & 1,456 & 2,480 & 1,364 & 1,910 & 1,622 \\ 1,522 & 0 & 1,568 & 2,014 & 1,399 & 1,239 & 2,274 & 1,420 & 1,725 & 1,099 \\ 1,847 & 1,568 & 0 & 2,243 & 1,673 & 1,792 & 2,012 & 1,778 & 0,964 & 1,729 \\ 1,936 & 2,014 & 2,243 & 0 & 2,243 & 1,575 & 2,789 & 1,847 & 2,352 & 1,667 \\ 1,735 & 1,399 & 1,673 & 2,243 & 0 & 1,661 & 1,881 & 1,493 & 1,836 & 1,723 \\ 1,456 & 1,239 & 1,792 & 1,575 & 1,661 & 0 & 2,375 & 1,382 & 1,967 & 1,063 \\ 2,480 & 2,274 & 2,012 & 2,789 & 1,881 & 2,375 & 0 & 2,332 & 2,121 & 2,304 \\ 1,364 & 1,420 & 1,778 & 1,847 & 1,493 & 1,382 & 2,332 & 0 & 1,873 & 1,526 \\ 1,910 & 1,725 & 0,964 & 2,352 & 1,836 & 1,967 & 2,121 & 1,873 & 0 & 1,921 \\ 1,622 & 1,099 & 1,729 & 1,667 & 1,723 & 1,063 & 2,304 & 1,526 & 1,921 & 0 \end{bmatrix}$$

As a comparison, the distances between Czechoslovakia and some other objects, calculated on the basis of complete data matrix are given as a vector:

$$d_2 = [1,396 \quad 0,000 \quad 1,470 \quad 1,987 \quad 1,285 \quad 1,269 \quad 2,100 \quad 1,296 \quad 1,673 \quad 1,304].$$

Then the distances between attributes were calculated. The distances between attributes X_2, X_3, X_4, X_5 and X_6 were determined on the basis of all data, the distances between X_1 and some other attributes on the basis of all data without the data concerning Czechoslovakia.

The distance matrix is as follows:

$$D = \begin{bmatrix} 0 & 0,120 & 0,588 & 0,790 & 0,846 & 0,668 \\ 0,120 & 0 & 0,721 & 0,633 & 0,780 & 0,551 \\ 0,588 & 0,721 & 0 & 0,645 & 0,986 & 0,667 \\ 0,790 & 0,633 & 0,645 & 0 & 0,602 & 0,157 \\ 0,846 & 0,780 & 0,986 & 0,602 & 0 & 0,430 \\ 0,668 & 0,551 & 0,667 & 0,157 & 0,430 & 0 \end{bmatrix}$$

As a comparison, the distances between X_1 and some other attributes (calculated on the basis of complete data matrix) are given as the following vector:

$$[I_1 = 0,000 \quad 0,120 \quad 0,601 \quad 0,790 \quad 0,846 \quad 0,668]$$

10. The authors' experience shows that the quality of data used in different fields of applied economics e.g. in econometric models, in statistical forecasting or financial reports is often

of poor quality. The gaps existing in the scope of data are filled up by some estimates obtained by means of various inter--or/and extrapolation techniques. The problem is extremaly important not only in econometric modelling but also in statistical multivariate international comparisons or in linear and nonlinear programming. In the present paper we aim at discussing the necessity of putting a stress on the data which are recommended as a statistical basis for the use in various economic analyses and experimental designs.

REFERENCES

[1] B a i l e r B. A. (1985), *Quality Issues in Measurements*, Inter. Statist. Rev., 153, 123-139.

Zdzisław Hellwig, Edward Nowak

PROBLEMY NIEDOSTATECZNEJ INFORMACJI
W MODELOWANIU TAKSONOMICZNYM

Celem artykułu jest prezentacja metod modelowania taksonomicznego w sytuacji brakujących informacji w zbiorze zmiennych objaśniających. Rozważa się sytuacje, w których brakujące informacje można uzupełnić, stosując metody wykorzystujące informacje zewnętrzne lub nie biorące pod uwagę tych informacji. W artykule proponuje się 2 metody wykorzystujące informacje zewnętrzne, a mianowicie:

- metoda zmodyfikowanej metryki Euklidesowej,
- metoda metryki korelacyjnej.