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 255, 2011

Andrzej Mantaj[*]

MULTIVARIATE ANALYSIS IN 3×3 TYPE TABLE

Abstract. In the paper there was presented the proposal of methods of analysis of features of spatial units in contingency tables on the example of four-fold table of 3×3 type. Apart from assessment of shaping of characteristics in the subclasses separated due to the adopted criterion features, there were applied the methods describing simultaneous shaping of variability in them, interdependences, and there were found influential spatial units, using generalized standard deviation, means from absolute values of differences between correlation coefficients, determinants of correlation matrixes, Kaiser-Meyer-Olkin statistics, feature values of correlation matrixes, and also projection matrixes and lever values.

Key words: statistics, four-fold tables, multi-feature observations.

I. INTRODUCTION

One of the methods enabling the analysis of properties of the set of observations simultaneously in respect of many features are, among other things, contingency tables, which also allow description of shaping the values of features of these observations within separated subclasses as more homogeneous subsets. For this purpose there are also used, belonging to classification methods, various hierarchic methods, methods of division and graphical presentation, and regression analysis (Gatnar and Walesiak 2004).

In this paper, for the analysis of tested units there were used four-fold tables of 3x3 type as a special case of contingency tables. The starting point to construct them is to determine the features constituting the criterion of division of observations, which determines the dimension of the table. In turn, the number of their subclasses is determined by the product of the assumed number of classes in respect to each of the criterion features according to the principle of Cartesian product. Simultaneous consideration of two or more features being the basis of division of observations allows, among other things, a general characteristic of the type of connections between values of these features and quantities of features describing the observations being in particular subclasses.

[*] dr, Wyższa Szkoła Informatyki i Zarządzania w Rzeszowie.

The traditional method of analysis of contingency tables focuses on calculation and interpretation of various numerical characteristics of observations in subclasses. In the paper there are presented the methods enabling description of properties of tested units both within subclasses as well as at simultaneous presentation of all values being in the four-fold table.

The purpose of the paper is the analysis of shaping of values of features of observations in the table created on the basis of values of two criterion features divided into three equinumerous classes each, i.e. in the 3×3 table

II. RESEARCH MATERIAL

The research material are the data of Statistical Office in Rzeszów, presented in the study „Województwo podkarpackie 2008 – podregiony, powiaty, gminy" *(Podkarpackie Province 2008 – subregions, districts, communes)*, concerning 143 rural and rural-town communes of Podkarpackie Province according to the state on 31.12.2007, and the quantities of the general quality coefficient of agricultural production space were determined on the basis of the study of the IUNG in Puławy (Witek 1993).

As a criterion of isolating the classes in the four-fold table of 3×3 type there were assumed tercentiles (quantiles of the rows 0,33 and 0,67) of the general quality coefficient of agricultural production space (QCAPS) and population density (PD). To the 9 subclasses separated in this way there were classified, in respect of assumed two criterion features, the communes which were characterized by 4 diagnostic features mentioned in table 1.

Table 1. Features describing the communes

Feature	Description	Designation
X_1	Balance of migration for 1000 persons	persons/1000 persons
X_2	Unemployment rate	%
X_3	Percentage of individual farms of area up to 3 ha UR	%
X_4	Incomes of communes in total for one dweller	thou. PLN /1 dweller

Source: The author's elaboration.

QCAPS is the sum of point assessment of 4 components expressed by indicators of valuation of quality and agricultural usability of soils, agricultural climate, relief and water conditions. For three communes, i.e. Besko, Gawłuszowice and Krościenko Wyżne, which were not considered in the study (Witek 1993), adequate values of the indicator were determined as arithmetic

mean of the indicators of neighbouring communes. The quantities of the remaining features presented in table 1 were calculated on the basis of the data of Statistical Office in Rzeszów, at the same time for the unemployment rate was assumed the quotient of the number of registered unemployed persons to the number of persons in productive age.

The numeration, assumed in the paper, of subclasses in the table of 3x3 type which was created on the basis of tercentiles of said criterion features QCAPS and PD, was presented in table 2.

Table 2. Specification of subclasses of 3 x 3 table

QCAPS	PD		
	1	2	3
3	(1,3)	(2,3)	(3,3)
2	(1,2)	(2,2)	(3,2)
1	(1,1)	(2,1)	(3,1)

Source: The author's elaboration.

All 143 communes were assigned to particular subclasses on the basis of values of their QCAPS and PD. It turned out that there were not many communes of the highest population density and the lowest valorisation indexes and the ones having the lowest population density and at the same time the highest valorisation indexes, and the quantities corresponding to these groups of communes in subclasses (1,3) and (3,1) were 4 and 5 respectively. It limited the possibilities of use of the assumed methods of analysis in the further part of the paper, and in this connection it was decided to incorporate these subclasses to one of neighbouring subclasses in relation to which they would show the greatest similarity. The basis of determining this similarity was the sum of absolute deviations of arithmetic means between subclasses for the values subjected to unitarization (Dziechciarz 2003). The values of unitarized means were calculated as:

$$\bar{z}_j = (\bar{x}_j - \min_i x_{ij}) / R_j, \tag{1}$$

where R_j is the range of the j-th diagnostic feature, and the distances between subclasses are expressed by:

$$d_{ikl} = \sum_{i=1}^{4} |\bar{z}_{ik} - \bar{z}_{il}|, \tag{2}$$

where i - number of feature (i = 1,...,4), k - subclass to which there was incorporated the observation, l – incorporated subclass. On the basis of the assumed procedure it turned out that the subclass (1,3) should be joined with subclass (1,2), and (3,1) with (2,2). The final number of communes in the 3×3 table was presented in table 3.

Table 3. Quantities of communes in joined subclasses

Population density	Number of communes			
	Valorisation index			Total
	< 65,1	65,1 - 76,4	> 76,4	
> 177,3	0	14	27	41
72,2 - 117,3	17	27	15	59
< 72,2	30	13	0	43
Total	47	54	42	143

Source: The author's elaboration.

The analysis of shaping the values of features of communes in the further part of the paper was conducted with consideration of division of communes in table 3.

III. PROPOSAL OF METHODS OF ANALYSIS OF MULTI-FEATURE DATA IN 3×3 TABLE

For the purpose of preliminary characterization of communes and shaping of the features describing them, there were calculated the values of arithmetic means in particular classes and subclasses of tested units, separated in respect of criterion features, i.e. QCAPS and PD.

As a measure of joint degree of diversification of values of all analysed features in subclasses and, for comparison, in the whole set of units, there was used the generalized standard deviation. It is the root of the degree determined by the number of variables from generalized variance being the determinant of variance-covariance matrix **S**. The elements of matrix $\mathbf{S} = (s_{ij})$, for $1 \le i, j \le p$ denote sample: variances $s_{jj} = s_j^2$ and covariances s_{ij}. It is assumed that the matrix **S** is positively determined, i.e. the determinants of all square submatrixes $\mathbf{S}_{kk}$ of the k-th degree, k = 1, ..., p created from the elements s_{ij} are positive. In par-

ticular the determinant $|\mathbf{S}| > 0$ and there occurs for it the inequality $|\mathbf{S}| \leq \prod_{j=1}^{p} s_{jj}$ (Fisz 1967).

The basis of assessment of diversification of interdependences of analyzed features in subclasses in respect to the central subclass (2,2) are the means of absolute values of differences between coefficients of matrixes $\mathbf{R} = (r_{ij})$ of Person's linear correlations in these subclasses.

$$\overline{d}_{(i,j)} = \frac{\sum_{k=2}^{p}\sum_{l=1}^{k-1}\left|r_{(i,j)_{kl}} - r_{(2,2)_{kl}}\right|}{p(p-1)/2}, \tag{3}$$

where: (*i, j*) - number of subclass, *i,j* = 1,2,3, (*2,2*) - central subclass , *p* - number of features, and *k, l* - feature numbers. Then from these means there was determined the percentage of differences of quantities smaller than the set threshold.

In order to determine the strength of joint connection of features in particular subclasses, there were calculated the determinants of matrixes of Pearson's correlation coefficients. The highest value equal to 1 is assumed by this determinant in the case of no correlation of variables, i.e. when in the correlation matrix beyond its main diagonal all elements are zeros, and at the linear dependence of any pair of features this determinant is zero (Pawłowski 1976).

For simultaneous assessment of the degree of connection between variables there was also applied Kaiser-Meyer-Olkin statistics (Malarska 2005), determined from the formula:

$$KMO = \frac{\sum_{j}\sum_{i\neq j} r_{ij}^2}{\sum_{j}\sum_{i\neq j} r_{ij}^2 + \sum_{j}\sum_{i\neq j} r_{ij.**}^2} \quad \text{dla } i, j = 1,2,...,p, \tag{4}$$

where r_{ij**} are partial correlation coefficients. The statistics *KMO* is the quotient of the sum of coefficients of determination between all pairs of variables and this quantity increased by the sum of squares of partial correlation coefficients. It assumes values in the interval [0, 1], indicating the gross relation of values of coefficients and the latter ones presented jointly with partial coefficients of determination.

On the basis of correlation matrixes of the analyzed variables within particular subclasses and for all communes jointly we also determined characteristic values, striving to determine the variance of each component, and at the same

time to determine the degree of explanation of total variability of variables taken into consideration in the analysis.

In order to find the diverging units among the tested communes there were determined their lever values with consideration of analysed features on the basis of ($n \times n$)-dimensional matrix of projection $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. It fulfils the property that its trace equals the number of features, i.e. $tr(\mathbf{H}) = p$. The lever values are defined as diagonal elements of matrix **H** and are contained in the interval $[1/n, 1]$. The commune was found diverging if the lever value corresponding to it exceeded the quantity $2p/n$, where p denotes the number of features, and n the number of tested communes.

IV. RESULTS OF ANALYSIS IN 3×3 TYPE TABLE

In table 4 there are presented the values of arithmetic means of analysed features describing the communes in individual subclasses and in classes separated in respect of criterion features, i.e. quality coefficient of production space and population density.

Table 4. Means of features in subclasses

Population density	Features	Quality coefficient of agricultural production space			
		< 65,1	65,1–76,4	> 76,4	Total
> 177,3	X_1		–0,16	–0,17	–0,17
	X_2		10,44	9,68	9,94
	X_3		70,21	73,03	72,07
	X_4		2,12	2,11	2,11
72,2 – 117,3	X_1	–0,37	0,58	–0,7	–0,03
	X_2	9,9	10,8	10,3	10,44
	X_3	52,6	57,2	54,09	55,06
	X_4	2,15	2,17	2,09	2,14
< 72,2	X_1	–1,8	–3,3		–2,3
	X_2	11,6	12,9		12,0
	X_3	41,4	42,1		41,6
	X_4	2,46	2,33		2,4
Total	X_1	–1,3	–0,6	–0,4	–0,7
	X_2	11,0	11,2	9,9	10,8
	X_3	45,4	56,9	66,3	55,9
	X_4	2,3	2,2	2,1	2,22

Source: The author's elaboration on the basis of data of Statistical Office in Rzeszów.

On the basis of table 4 it is observed that together with the increase of QCAPS in the whole set of communes there grew population density and the percentage of farms of area up to 3ha (X_3), and there decreased the unemploy-

ment rate (X_2), general income of communes per capita (X_4) and there decreased the degree of migration (X_1). Taking into consideration the fact that there exists the positive correlation between QCAPS and PD, which is also indicated by the distribution of quantities of communes in subclasses (table 3), to the increase of population number in reference to the unit of surface there correspond dependences similar to the previous ones, except poorer connection with the indicator of migration (X_1). These connections within the separated subclasses are confirmed in the majority of cases, however, sometimes they are less distinct, as, for example, for the indicator of migration (X_1) and the unemployment rate (X_3).

The joint degree of diversification of values of analysed features in subclasses and in the whole set of units is illustrated by the values of generalized standard deviation presented in table 5.

Table 5. Generalized standard deviation in subclasses

PD	QCAPS		
	< 65,1	65,1–76,4	> 76,4
> 177,3		0,47	0,94
72,2 – 117,3	0,36	0,52	0,31
< 72,2	0,76	0,64	
Total	1,02		

Source: The author's elaboration.

The highest quantity of the generalized deviation (0,94) was obtained in the subclass (3,3), at the same time it was smaller than for the whole set of units, equal 1,02. The communes in subclasses (3,2) and (1,2) turned out to be the most similar to each other.

The degree of similarity of analysed features in subclasses relative to the central subclass (2,2), expressed by the percentage of means of absolute values of differences between Person's correlation coefficients in these subclasses exceeding the set value equal 0,15 is presented in table 6.

Table 6. Percentage of means of differences

PD	QCAPS		
	< 65,1	65,1 - 76,4	> 76,4
> 177,3		66,67	50,00
72,2 - 117,3	33,33	×	66,67
< 72,2	16,67	100,00	
Total	66,67		

Source: The author's elaboration.

The most similar, in respect of quantities of correlation coefficients, to the central subclass was the subclass (2,1), and the most diverging from the central subclass was the subclass (1,1). The determinants of Pearson's correlation matrix, as measures of strength of joint connection of features in individual subclasses, are given in table 7.

Table 7. Determinants of Pearson's correlation matrix

PD	QCAPS		
	< 65,1	65,1 - 76,4	> 76,4
> 177,3		0,74	0,75
72,2 - 117,3	0,52	0,85	0,51
< 72,2	0,40	0,74	
Total	0,76		

Source: The author's elaboration.

The highest degree of correlation of features (0,40) is found in the subclass (1, 1), so in the communes of the lowest values of QCAPS and PD, and the poorest correlation of features (0,85) occurred in the central subclass. For simultaneous assessment of the degree of connection between features there was also applied Kaiser-Meyer-Olkin statistics whose values are presented in table 8.

Table 8. Kaiser-Meyer-Olkin statistics

PD	QCAPS		
	< 65,1	65,1–76,4	> 76,4
> 177,3		0,48	0,38
72,2–117,3	0,47	0,50	0,33
< 72,2	0,54	0,48	
Total	0,62		

Source: The author's elaboration

The connections between the tested features, not taken into consideration in these coefficients in the subclasses (3,2) and (3,3), had high part in gross correlation coefficients. In the remaining classes the relations were similar and oscillated around 0,5.

The degree of explanation of total variability of features taken into consideration in the analysis, on the basis of two first characteristic values, is presented in table 9.

Table 9. First and second characteristic values and cumulated percentages of variability in subclasses

PD	Quality coefficient of agricultural production space						
	Components	< 65,1		65,1–76,4		> 76,4	
		1	2	1	2	1	2
> 177,3	Characteristic values			1,40	1,33	1,40	1,12
	Cumulated. % of variance			35,06	68,26	34,90	63,02
72,2–117,3	Characteristic values	1,56	1,45	1,42	1,04	1,64	1,16
	Cumulated. % of variance	39,10	75,42	35,59	61,64	40,95	70,06
< 72,2	Characteristic values	1,95	1,20	1,48	1,22		
	Cumulated. % of variance	48,71	78,65	36,89	67,36		
Total	Characteristic values	1,66, 0,93					
	Cumulated. % of variance	41,43, 64,59					

Source: The author's elaboration.

In all subclasses only the two first components assume values greater than one, and in the whole set the first component explains over 40% of total variability. The cumulated value of variance explained (78,65%) by two first components is highest in the subclass (1,1).

Lever values with consideration of analyzed features allowed finding the most diverging communes which are presented in table 10.

Table 10. Influential communes in subclasses

PD	QCAPS		
	< 65,1	65,176,4	> 76,4
> 177,3		-	Świlcza Trzebownisko
72,2117,3	Głogów Młp. Rymanów	Domaradz Wojaszówka	Lubenia Wiśniowa
< 72,2	Pysznica Wielkie Oczy	-	
Total	Cieszanów, Domaradz, Stary Dzików, Świlcza, Tyczyn and Wiśniowa		

Source: The author's elaboration.

The communes diverging from the remaining ones and the values differing clearly from the means were not observed only in subclasses (2,1) and (2,3). In the remaining subclasses there occurred always by two influential communes.

For the purpose of identifying the number of features whose values could influence diverging of communes from the remaining ones, in table 11 there are presented influential communes with marking which of the features describing them were close or equal to extreme quantities.

Table 11. Max and min values of features in influential communes

Subclasses	Commune	Features			
		X_1	X_2	X_3	X_4
(1,1)	Pysznica	max	min	max	
	Wielkie Oczy			≈ min	≈ min
(1,2)	Głogów Młp.	max	≈ min		min
	Rymanów		min	≈ max	
(2,2)	Domaradz		max	≈ max	
	Wojaszówka		min	≈ max	min
(3,2)	Lubenia			max	min
	Wiśniowa		max		
(3,3)	Świlcza	min			
	Trzebownisko				max
Total	Cieszanów			≈ min	
	Domaradz			max	
	Pysznica			min	
	Stary Dzików		≈ min	min	
	Świlcza	min			
	Trzebownisko	≈ sr	≈ sr	≈sr	
	Tyczyn	max			
	Wiśniowa		≈ min		

Source: The author's elaboration.

The value of at least one feature for communes identified as diverging in particular subclasses is the extreme value, at the same time 3 communes have one or three extreme values, and 5 communes – two such values. For 8 diverging communes found in the whole set and 9 selected in subclasses, 5 of them are repeated, i.e. they were found diverging from the remaining ones regardless of the type of set in which there were determined. Only in one case as diverging was found the commune showing quantities of features close to the means, and finding it influential could be caused by different system of feature values for this commune in comparison with the other ones.

V. SUMMARY

The analysis of observations in contingency tables, which was demonstrated on the example of four-fold table of 3×3 type, can provide much interesting information on the analysed set of spatial units. It allows not only the possibility of observing the tendencies of shaping of characteristics of different type in the subclasses separated in respect of the assumed criterion features, but also using methods describing simultaneous shaping of variability and interdependence in them and finding influential spatial units. For this purpose there are used such measures as generalized standard deviation, means of absolute values of differences between correlation coefficients, determinants of correlation matrixes, Kaiser-Meyer-Olkin statistics, characteristic values of correlation matrixes, and also matrixes of projection and lever values.

BIBLIOGRAPHY

Dziechciarz J. (2003), *Ekonometria (Econometrics)*, AE Wrocław.
Gatnar E., Walesiak M. (2004), *Metody statystycznej analizy wielowymiarowej w badaniach marketingowych (Methods of Statistical Multivariate Analysis in Marketing Tests)*, AE Wrocław.
Fisz M. (1967), *Rachunek prawdopodobieństwa i statystyka matematyczna (Calculus of Probability and Mathematical Statistics)*, PWN, Warszawa.
Malarska A. (2005), *Statystyczna analiza danych wspomagana programem SPSS (Statistical Analysis of Data Enhanced by SPSS Program)*, SPSS Polska, Kraków.
Pawłowski Z. (1976), *Statystyka matematyczna (Mathematical Statistics)*, PWN, Warszawa.
Witek T. (1993), *Waloryzacja rolniczej przestrzeni produkcyjnej Polski według gmin (Valorisation of Agricultural Production Space of Poland According to Communes),* Puławy.

Andrzej Mantaj

ANALIZA WIELOWYMIAROWA W TABLICY TYPU 3×3

Praca zawiera propozycję metod analizy cech jednostek przestrzennych w tabelach wielodzielczych, przedstawioną na przykładzie tabeli dwudzielczej typu 3 x 3. Do prezentacji zastosowanych metod wykorzystano dane dotyczące 143 gmin wiejskich i miejsko-wiejskich woj. podkarpackiego według stanu na 31.12.2007 r.

Przedstawiona w pracy analiza, poza oceną kształtowania się charakterystyk w wydzielonych ze względu na przyjęte cechy kryterialne podklasach, objęła również przykłady zastosowania metod jednoczesnego opisu zmienności i współzależności badanych charakterystyk oraz ustalanie wpływowych jednostek przestrzennych, wykorzystując w tym celu uogólnione odchylenie standardowe, średnie z bezwzględnych wartości różnic między współczynnikami korelacji, wyznaczniki macierzy korelacji, statystykę Kaisera-Meyera-Olkina, wartości własne macierzy korelacji, a także macierze rzutowania i wartości dźwigniowe.