ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 235, 2010

Andrzej Mantaj[*], *Robert Pater*[**], *Wiesław Wagner*[***]

ASPECTS OF LINEAR AND MEDIAN CORRELATION COEFFICIENTS MATRIX

Abstract. Linear and correlation coefficients characterized in this paper indicate the usefulness in analyzing the relation between their qualities, especially for big samples. The first of them is recommended in situation where two-dimensional samples do not diverge to the direction of each of coordinate axis in correlative graph so two-dimensional diverging observations do not occur. Their occurrence leads to arithmetic means refraction what marks classical Pearson's correlation coefficient as low resistant statistics describing quality of dependency. These circumstances are described in this paper though analyzing number material.

Median correlation coefficient recommended for big samples shows high resistance to diverging observations. It is characterized by many qualities which are analogical to linear correlation coefficient.

Median and linear correlation coefficient can be used in case of two-dimensional samples coming from population of two-dimensional normal dispersion.

Next to presented correlation coefficients, Spearman's rang correlation coefficient is used in practice as well. It is the proposition of examining dependency of qualities measured in subordinated scale. Here the distance between two-dimensional sample observations is not taken into account, but rather the order of qualities when they are measured in smaller scales .

Key words: linear correlation coefficient, median correlation coefficient.

I. INTRODUCTION

The linear correlation coefficients matrix plays a fundamental role in examining the dependency of given system values in multidimensional statistical analysis. The matrix lets determine the correlated values on the basis of diagonal values of its inverse matrix as well as interpretation of variables by using number statistics charts and graphs created on its elements. By constructing the matrix classical number characteristics of position and variables are used.

The alternative of mentioned matrix is median correlation matrix. Median correlation matrix is a proposition for analyzing the dependency of qualities set, where some of them reveal the existence of divergent observations found quite

[*] Ph. D., University of Information Technology and Management in Rzeszów.

[**] MA, University of Information Technology and Management in Rzeszów.

[***] Professor, University of Information Technology and Management in Rzeszów.

often in analysis of social-economic phenomena. Then some difficulties with quality evaluation occur. The phenomenon has been caused by low value of refraction of arithmetic mean (Rousseeuw and Lorey 1987). So instead of arithmetic mean, marked as a classical characteristic of number position, median, positional characteristic of number is applied and characterized by high value of refraction. Due to this, the pair {arithmetic mean, standard deviation} in linear correlation is substituted by the pair {median, median absolute deviation} in median correlation.

The aim of this paper is to compare linear and median correlation coefficients in empirical material. Different formulas of determining linear correlation coefficients have been provided and the basic values of median absolute deviations of one-dimensional sample and comedian for two- dimensional sample were described. The last two are used for calculating median correlation coefficient, which values are demonstrated in this paper as well.

The research material consists of number data of 14 social- economic diagnostic values for rural districts and urban- rural districts in Podkarpacie providence for the day of 31.12.2002

II. RESEARCH MATERIAL

The research material constitutes data from statistical yearbook 2004 of Podkarpacie providence (Statistical Office in Rzeszow). The data include 144 rural and urban-rural districts of Podkarpackie providence for the day of 31.12.2002 taking into consideration Szerzyna district, which was excluded from Jasielski area in 2003. The list of 14 variables of social-economic diagnostic for districts is given in table 1.

Table 1. Diagnostic values of urban- rural districts

Variables	Description	Symbols
X_1	population density	Population/ km2
X_2	population in productive age	%
X_3	birth rate per 1000 people	-
X_4	migration balance per 1000 people	-
X_5	employment rate (factual number of the employed)	%
X_6	unemployment rate (comparison of the registered unemployed to the employed)	-
X_7	average floor surface in m2 per 1 inhabitant	m^2/ 1 inhabitant
X_8	comparison of farms over 3 ha to farms below 1 ha	%
X_9	relation between arable land surface to 3 ha and agricultural land over 1 ha	%
X_{10}	forestry	%
X_{11}	shop number per 1000 inhabitants	-
X_{12}	total incomes per 1 inhabitant (zl)	zł. (Polish zloty)
X_{13}	individual income to income in general	-
X_{14}	number of registered national economy subjects per 100 people in productive age	% 100 people

Source: own elaboration on the basis of Statistical Yearbook of Podkarpackie Province, 2004.

Data number presented in table 1 were earlier used in Mantaj and Wagner thesis (2007) and they are available at amantaj@wsiz.rzeszow.pl

III. LINEAR CORRELATION COEFFICIENT

Let there be two-dimensional sample of *n*-element observations for (*X, Y*) values, expressed by set of *n*-pairs $(x_1, y_1),...,(x_n, y_n)$ or a pair of two *n*-dimensional column vectors (**x**, **y**). The sample of these observations will be indicated by P_n^2. The sample linear correlation coefficient $r = r(P_n^2)$ of P_n^2 sample can be indicated in many ways, using:

a) sum of squares deviated and sum cross products deviated

$$r = \frac{SXY}{\sqrt{SSX \cdot SSY}} \tag{1}$$

where:

$$SSX = \sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i(x_i - \overline{x}) = \sum_{i=1}^{n} x_i^2 - n \cdot \overline{x}^2 = \mathbf{x'Ax} = \mathbf{x'x} - \mathbf{x'Jx},$$

$$SSY = \sum_{i=1}^{n}(y - \overline{y})^2 = \mathbf{y'Ay} = \mathbf{y'y} - \mathbf{y'Jy},$$

$$SXY = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{n} x_i(y_i - \overline{y}) = \sum_{i=1}^{n}(x_i - \overline{x})y_i = \sum_{i=1}^{n} x_i y_i - n \cdot \overline{xy} =$$
$$= \mathbf{x'Ay} = \mathbf{x'y} - \mathbf{x'Jy},$$

are appropriately the sums of deviation squares of average *X,Y* values and the sum of deviation products of average for *X,Y* values, where matrix **A** = **I** – **J**, whereas $\mathbf{J} = \mathbf{11'}/n$.

b) sample variances and covariances

$$r = \frac{s_{xy}}{s_x s_y} \tag{2}$$

where $s_x = \frac{SSX}{n-1}$, $s_y = \frac{SSY}{n-1}$, $s_{xy} = \frac{SXY}{n-1}$ are sample variances and covariance of *X, Y* values;

(c) normalized observations of *X, Y* variables

$$r = \sum_{i=1}^{n} (t_i - \bar{t})(z_i - \bar{z}), \tag{3}$$

where $t_i = \frac{x_i}{\sqrt{SSX}}$, $z_i = \frac{y_i}{\sqrt{SSY}}$;

(d) standardized sample of observation variables

$$r = \frac{1}{n-1} \sum_{i=1}^{n} u_i v_i, \tag{4}$$

where $u_i = \frac{x_i - \bar{x}}{s_x}$, $v_i = \frac{y_i - \bar{y}}{s_y}$.

Formula (1) is regarded as primal in indicating linear correlation coefficient, whereas formulas (2), (3) and (4) are of different versions. By determining correlation coefficient

(i) $r = r(x, y)$ as *X, Y* values, qualities can be added,

(ii) $r = r(x, y) = r(y, x)$- symmetry property,

(iii) $r(x, x) = 1$ – reflexivity property,

(iv) $r \in \langle -1, 1 \rangle$, because $|SXY| \le \sqrt{SSX \cdot SSY}$,

(v) $r(ax + b, cy + d) = r(x, y)$) – property of invariability on account of linear transformation of each value.

Formula (1) is recognized as a formula of general correlation coefficient, where central moments are substituted by centered moments of *a, b* constant

$$r = \frac{\sum_{i-1}^{n} (x_i - a)(y_i - b)}{\sqrt{\sum_{i=1}^{n} (x_i - a)^2 \sum_{i=1}^{n} (y_i - b)^2}} \tag{5}$$

By replacing *a, b* constant with the position of different number characteristics, then various formulas of correlation coefficients are obtained. It is applied to such characteristics as Hogdes mean (median of all average pairs of observations), median, average cuts, quantile including upper quartile and lower quartile

Example 1. For two-dimensional sample

Number	1	2	3	4	5	6	7	8	9	10
x	9	14	13	17	21	22	25	27	29	34
y	32	38	42	47	46	49	50	51	54	62

correlation coefficient is computed by (5) formula using various *a, b:*

Statistics	a	b	Correlation
Arithmetic means	21,1	47,1	0,9601
Hodges means	21,5	48	0,9586
Median	21,0	47,5	0,9581
Lower quartile	14,75	43,0	0,9468
Upper quartile	26,5	50,75	0,9514

The defined values of correlation coefficients do not differed from one another. The highest value was obtained for *a, b* constants adequately to values of arithmetic means. Let there be indicated that by determining linear correlation coefficient (1) the basic operation of number values is to sum square deviations and product deviations of arithmetic means. This model applies to notions connected with defining normal and central moments for random variables of jump-types. The calculation of correlation coefficient requires defining arithmetic means for both values.

Example 2. For data in example 1 a couple of correlation coefficients will be defined, when the last pair of observations is gradually changed. The results are presented below:

X changing values		Y changing values	
Pair of observation	Correlation	Pair of observation	Correlation
(45, 62)	0,9466	(34, 78)	0,8986
(55, 62)	0,9074	(34, 88)	0,8543
(75, 62)	0,8339	(34, 95)	0,8278

In every case the values of one quality go up and the values of correlation coefficient go down.

IV. LINEAR CORRELATION COEFFICIENT MATRIX

For $P_n^p = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n),$ multidimensional sample *n* containing n vectors of observations of *p*-dimensional, linear correlation coefficients matrix sample are indicated as in $\mathbf{R} = (r_{jk})$, *j, k* = 1, 2, ..., *p*. Matrix's elements are indicated by

$r_{jk} = \dfrac{s_{jk}}{\sqrt{s_{jj} \cdot s_{kk}}}$, where quantities in numerator and denominator are elements of variance and covariance sample matrix $\mathbf{S} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})' = (s_{jk})$, and is p-dimensional vector of values of arithmetic means. s_{jj}, s_{kk}, $j \neq k$ elements are diagonal ones and s_{jk} out of diagonal **S** matrix. For **R** matrix elements cohesive graph is formulated, according to:

a) elements r_{jk} in $d_{jk} = 1 - r_{jk}^2$ are transformed, so low values of distance measure of d_{jk} are adequate with high values of correlation coefficients,

b) minimal values of d_{jk} are indicated for every quality,

c) quantity node connections of the closest d_{jk} values are performed by systematic deleting of values that are added to the graph.

Example 3. The of linear correlation coefficient matrix will be conducted on the basis of empirical data and presented in chapter two. The correlation matrix for 14 variables is presented in the table below:

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}
X_1	1,000													
X_2	0,219	1,000												
X_3	-0,023	0,191	1,000											
X_4	0,249	0,020	-0,132	1,000										
X_5	0,210	0,490	0,111	-0,102	1,000									
X_6	0,310	0,418	0,044	0,073	0,898	1,000								
X_7	0,141	-0,296	-0,149	0,186	-0,390	-0,154	1,000							
X_8	0,755	0,268	-0,093	0,331	0,138	0,185	-0,071	1,000						
X_9	0,324	-0,417	-0,151	0,046	-0,151	-0,003	0,363	0,145	1,000					
X_{10}	-0,663	0,235	0,095	-0,139	0,086	-0,108	-0,409	-0,380	-0,639	1,000				
X_{11}	-0,029	0,495	0,051	-0,223	0,457	0,327	-0,419	-0,025	-0,426	0,340	1,000			
X_{12}	-0,329	0,122	0,162	-0,149	0,070	-0,040	-0,203	-0,251	-0,373	0,326	0,104	1,000		
X_{13}	0,195	0,561	0,108	0,112	0,470	0,391	-0,206	0,229	-0,184	0,074	0,283	0,471	1,000	
X_{14}	0,060	0,667	0,148	0,109	0,368	0,293	-0,385	0,205	-0,611	0,363	0,595	0,287	0,510	1,000

Source: Own elaboration.

The matrix constitutes the base of different interpretations of values of linear correlation:

(a) Analyze of correlation coefficients quantity:

- high negative values for X_2, X_9 (-0,417), X_7, X_{10} (-0,419), X_9, X_{10} (-0,611) and X_9, X_{10} (-0,639),

- high positive values for X_2, X_{14} (0,667), X_1, X_8 (0,755), X_5, X_6 (0,898),
- maximum range of correlation coefficient 0,898 – (-0,639) = 1,537 between pairs (X_9, X_{10}) and (X_5, X_6),
- the highest absolute value of negative (-) and positive(+) correlation coefficient with their ranges (R) and percentage shares in maximum gap (%) for separate values are in table 2

Table 2. the highest positive (+) and negative (-) correlations and their ranges

Variables	X_1	X_2	X_3	X_4	X_5	X_6	X_7
(-)	X10 -0,663	X9 -0,417	X9 -0,151	X11 -0,223	X7 -0,390	X7 -0,154	X11 -0,419
(+)	X8 0,755	X14 0,667	X2 0,191	X8 0,331	X6 0,898	X5 0,898	X5 0,363
R %	1,418 90,83	1,084 69,41	0,341 21,86	0,553 35,45	1,288 82,49	1,052 67,37	0,782 50,08
Variables	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}
(-)	X10 -0,380	X10 -0,639	X1 -0,663	X7 -0,426	X12 -0,373	X7 -0,206	X9 -0,611
(+)	X1 0,755	X7 0,363	X14 0,363	X14 0,595	X13 0,471	X2 0,561	X2 0,667
R %	1,135 72,67	1,002 64,20	1,026 65,72	1,021 65,41	0,844 54,08	0,768 49,16	1,278 81,85

Source: Own elaboration.

- the lowest ranges for correlations have the following qualities X_3,X_9 (21,86 %) and X_4, X_{11} (35,45 %), and the highest: X_{14}, X_9 (81,85 %), X_5, X_7 (82,49 %) and X_1,X_{10} (90,83 %),
- non diminishing ordered of correlation coefficients of separate qualities with marked range of negative correlations (grey colored) and with average changes has been shown in the list:

Variables	Numbers of variables in order of increasing correlation qualities													Average changes
1	10	12	11	3	14	7	13	5	2	4	6	9	8	0,128
2	9	7	4	12	3	1	10	8	6	5	11	13	14	0,109
3	9	7	4	8	1	6	11	10	13	5	14	12	2	0,089
4	11	12	10	3	5	2	9	6	14	13	7	1	8	0,094
5	7	9	4	12	10	3	8	1	14	11	13	2	6	0,107
6	7	10	12	9	3	4	8	14	1	11	13	2	5	0,089
7	11	10	5	14	2	13	12	6	3	8	1	4	9	0,109
8	10	12	3	7	11	5	9	6	14	13	2	4	1	0,106

9	10	14	11	2	12	13	3	5	6	4	8	1	7	0,126
10	1	9	7	8	4	6	13	5	3	2	12	11	14	0,128
11	9	7	4	1	8	3	12	13	6	10	5	2	14	0,110
12	9	1	8	7	4	6	5	11	2	3	14	10	13	0,106
13	7	9	10	3	4	1	8	11	6	5	12	14	2	0,093
14	9	7	1	4	3	8	12	6	10	5	13	11	2	0,124

Source: Own elaboration

- average changes of correlation coefficients were indicated out of earlier list, as average of 13 absolute increase between two adjacent correlations in the row and had quantities from 0,089 for X_6 quality to 0,128 for X_1 and X_{10} quality,
- for 168 (=14 ·12) increases got min = 0 and max = 0,809,
- juxtaposition of all increases presented in histogram of population, where *x*-axis present upper limit of class range (fig.1)

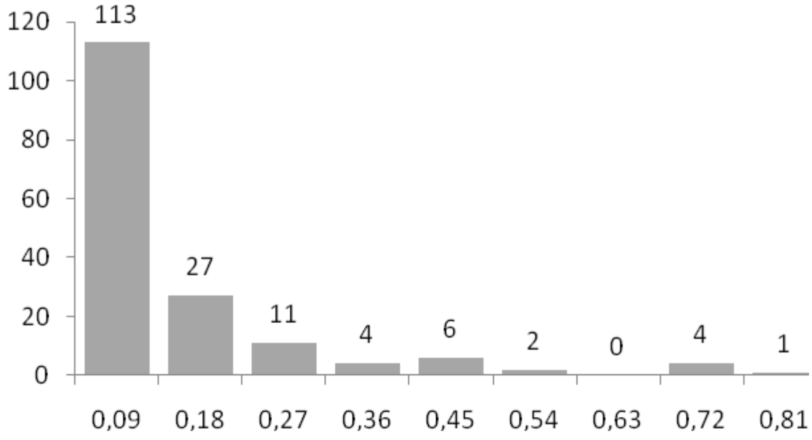

Fig.1. Population histogram of adjacent correlations increases

Source: Own elaboration.

- Linear correlation graph of X_1, X_2 quality in pairs with X_3, X_4, …., X_{14} (fig.2)

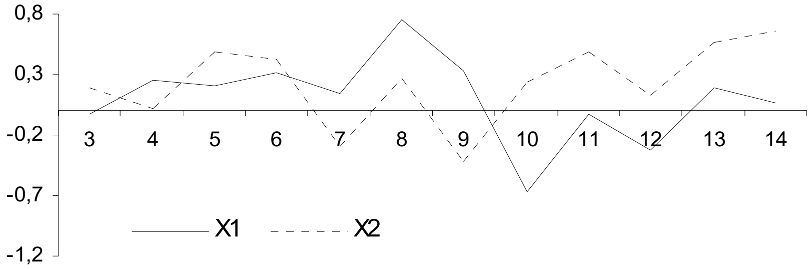

Fig.2 linear correlation graph of X_1, X_2 qualities with the rest qualities
Source: Own elaboration.

(b) correlative graph analysis:

The strongest negative correlation appeared in X_9, X_{10} quality (fig.3) and the strongest positive correlation appeared in X_5, X_6 quality (fig. 4)

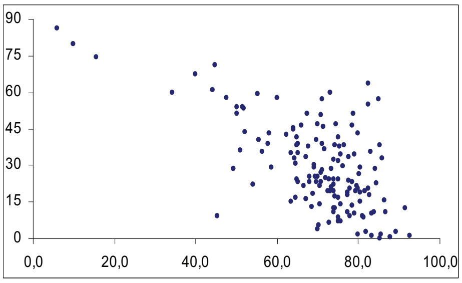

Fig.3 X_9 and X_{10} correlation qualities-
Source: Own elaboration

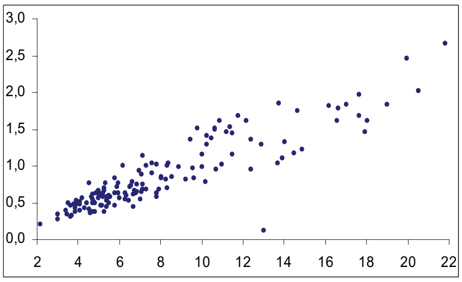

Fig.4. X_5 and X_{10} correlation qualities
Source: Own elaboration

(c) class rows and graphs:

- distribution line and population histogram (fig.5)

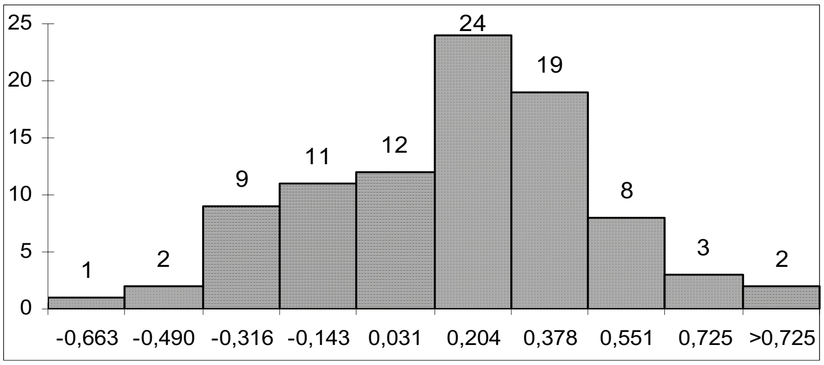

Fig.5. Population histogram of linear correlation coefficients
Source: Own elaboration.

- distribution line with apriori range of constant length 0,1 for correlation coefficients of differentiated pairs

-0,7	-0,6	1	9	9												
		10	10	14												
-0,5	-0,4	9	7	2	7											
		11	11	9	10											
-0,4	-0,3	5	7	8	9	1										
		7	14	10	12	12										
-0,3	-0,2	2	8	4	7	7										
		7	12	11	13	12										
-0,2	-0,1	9	6	3	5	3	4	4	3	6	4					
		13	7	9	9	7	12	10	4	10	5					
-0,1	0,0	3	7	6	1	8	1	6								
		8	8	12	11	11	3	9								
0,0	0,1	2	3	4	3	1	5	4	10	5	3					
		4	6	9	11	14	12	6	13	10	10					
0,1	0,2	11	3	4	3	4	2	5	1	8	3	3	6	4	2	1
		12	13	14	5	13	12	8	7	9	14	12	8	7	3	13
0,2	0,3	8	1	1	8	2	1	2	11	12	6					
		14	5	2	13	10	4	8	13	14	14					
0,3	0,4	1	1	10	6	4	10	7	10	5	6					
		6	9	12	11	8	11	9	14	14	13					
0,4	0,5	2	5	5	12	2	2									
		6	11	13	13	5	11									
0,5	0,6	13	2	11												
		14	13	14												
0,6	0,7	2														
		14														
0,7	0,8	1														
		8														
0,8	0,9	5														
		6														

Source: Own elaboration

(d) bar graphs of correlation coefficients for a distinguished quality with the remaining quality of higher numbers (fig. 6,7,8,9- qualities: X_1, X_2, X_3, X_4)

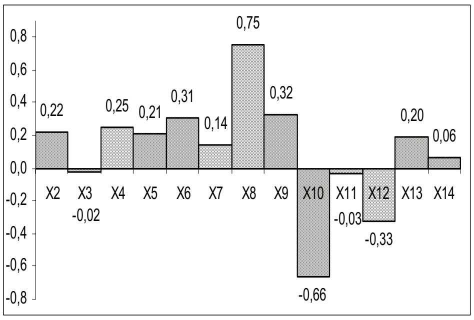

Fig. 6. X_1 quality and the remaining
Source: Own elaboration

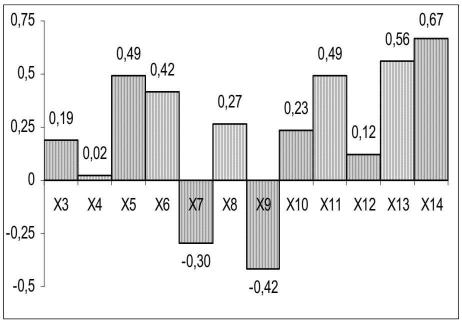

Fig. 7. X_2 quality and the remaining
Source: Own elaboration

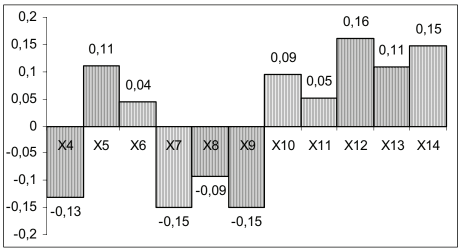

Fig. 8. X_3 quality and the remaining
Source: Own elaboration

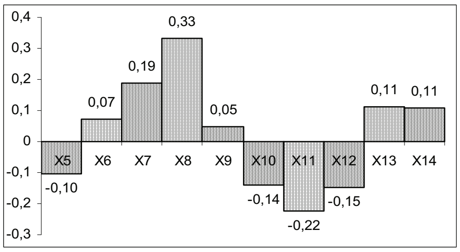

Fig. 9. X_4 quality and the remaining
Source: Own elaboration

(e) diagonal elements of inverse matrix to correlation matrix

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}
4,855	2,684	1,140	1,578	10,982	9,002	1,918	3,145	2,662	3,586	2,197	2,101	2,814	3,443

(f) cohesive graph (fig.10)

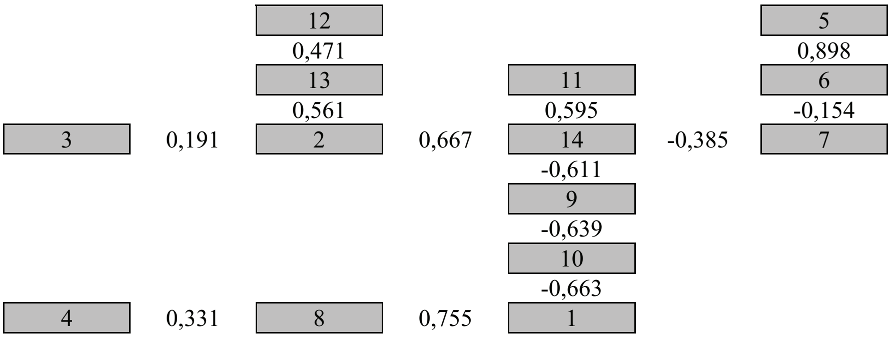

Fig.10. Cohesive graph of linear correlation coefficients
Source: Own elaboration

(g) binary matrix (two- elements) and ternary matrix (three- elements)

- replacing ternary correlation matrix with -1,0,1 elements for three distinguished ranges of linear correlation values, adequately -1 for correlation of <-1, *a)* range, 0 for <*a,b)* range and 1 for <*b,*1) range, where *a*= -0,2, b= 0,2

	1	2	3	4	5	6	7	8	9	10	11	12	13
2	1												
3	0	0											
4	1	0	0										
5	1	1	0	0									
6	1	1	0	0	1								
7	0	-1	0	0	-1	0							
8	1	1	0	1	0	0	0						
9	1	-1	0	0	0	0	1	0					
10	-1	1	0	0	0	0	-1	-1	-1				
11	0	1	0	-1	1	1	-1	0	-1	1			
12	-1	0	0	0	0	0	-1	-1	-1	1	0		
13	0	1	0	0	1	1	-1	1	0	0	1	1	
14	0	1	0	0	1	1	-1	1	-1	1	1	1	1

Source: Own elaboration

- other propositions in creating ternary matrix are::

❖ -1 for $\langle -1, \bar{r} - s_r))$ range, 0 for $\langle \bar{r} - s_r, \bar{r} + s_r)$ and 1 for $\langle \bar{r} + s_r, 1 \rangle$, where $\bar{r}, s_r$ are arithmetic mean and standard deviation of correlation coefficients,

❖ -1 for $\langle -1, Me_r - Mad_r))$ range, 0 for $\langle Me_r - Mad_r, Me_r + Mad_r)$ and *1* for $\langle Me_r + Mad_r, 1 \rangle$, where Me_r, Mad_r are median and median absolute deviation of correlation coefficients.

IV. MEDIAN CORRELATION COEFFICIENT

The linear correlation coefficient belongs to classical sample of two-dimensional number characteristics. Its alternative is median correlation coefficient belonging to positional sample of two-dimensional number characteristics. It is determined by use of median. In this case sample variability of positional characteristic is introduced, known as median absolute deviation and indicated by Mad (median absolute deviation). For sample of one-dimensional

$P_n^1 = \{x_1, x_2, ..., x_n\}$ is:

$$Mad = Med_i(|x_i - Med_j(x_j)|). \tag{6}$$

It is noticed from formula (6) that they are determined as median from absolute deviation from median. It is an equivalent of standard deviation being classical characteristic of variability. Mad(x) as a sample function P_n^1 fulfils conditions:

(a) $Mad(a) = 0$, for any *a* constant,

(b)$Mad(ax + b) = |a|Mad(x)$, for any *a, b* constants,

(c)$Mad(x) = Med(x) - \ Q_1(x) = Q_3(x) - Med(x)$ for X of symmetrical dispersion,

(d)$Mad(x - \bar{x}) = Q_3(x)$, observation of deviation from average sample of symmetrical dispersion,

(e)$Mad(x) = \sigma\Phi^{-1}(0{,}75)$ (Falk 1997), when the sample comes from population of normal dispersion $N(0, \sigma)$, but $\Phi^{-1}(.)$ is reverse distribution of dispersion distribution *N*(0, 1),

Case (c) and (d) are illustrated in the example. A $n = 7$ sample is given for observations ordered increasingly: 2*, 5, 9, 12, 15, 19, 22. Med.* = *12* is median in sample. Deviations of observations are the same around median, so $Med - x_i = x_{8-i} - Med$ for *i* = *1, 2 ,3* what means that a given sample has symmetrical dispersion. $| x_i - Med |$ absolute deviation is calculated and equals: *10, 7, 3, 0, 3, 7, 10*. They constitute the ordered sample : *0, 3 ,3, 7, 7, 10, 10* and *Mad* = *7.* Let's notice that for primal sample we have $Q_1 = 5$ and $Q_3 = 19$ that means $Med - Q_1 = 7$ and $Q_3 - Med = 7$. In condition (d) it is noticed that through analyzing arithmetic mean covers with median, so we have deviation : -10, -7, -3,0, 3,7, 10, where quantile three is equals *7* and covers with Mad.

Example 4. A $n = 7$ sample is given for observations ordered increasingly: 2, 5 , 9, 12, 15, 19, 22. *Med=12* is median in sample. Deviations of observations are the same around median, so $Med - \ x_i = x_{8-i} - Med$ for *i* = 1, 2 ,3 what means that a given sample has symmetrical dispersion.$|x_i - Med|$, absolute deviation is calculated and equals: 10, 7, 3, 0, 3, 7, 10. They constitute the ordered sample : 0, 3, 3, 7, 7, 10, 10 and *Mad* = *7*. Let's notice that for primal sample we have Q_1 = 5 and Q_3 = 19 that means $Med - Q_1 = 7$ and $Q_3 - Med = 7$. In condition (d) it is noticed that through analyzing arithmetic mean covers with median, so we have deviation: *-10, -7, -3,0, 3,7, 10*, where quantile three is equals *7* and covers with *Mad.*

In formula (2) denominator will be replaced by Mad(x) with Mad(y). For nominator is concerned statistic called comedian of P_n^2 sample is taken, what is written down as CoMed. It is constituted as a formula (Falk 1997, 1998; Wagner 1998 et. al.):

$$CoMed = CoMed(x,y) = Med_i\left\{\left(x_i - Med_j(x_j)\right)\left(y_i - Med_j(y_j)\right)\right\} \quad . \qquad (7)$$

For CoMed statistics the conditions are:

(a) $CoMed(x,x) = Mad^2(x)$,

Proof.

$CoMed(x,x) = Med\{(x - Med(x))(x - Med(x))\} = Med\{(x - Med(x))^2\}$ $= Med\{|x - Med(x)|^2\} = [Med\{|x - Med(x)|\}]^2 = Mad^2(x)$, with $Med(x^2) = [Med(x)]^2$, why $x > 0$;

(b)$CoMed(x,ax + b) = aMad^2(x)$,

(c) $CoMed(ax + b, cy + d) = ac\ CoMed(x\ ,y)$,

(d)$CoMed(x,y) = Med(x \cdot y) - Med(x) \cdot Med(y)$, out of which X,Y values are independent so $CoMed(x,y) = 0$,

(e)$CoMed(x,y) \le Mad(x) \cdot Mad(y)$, if P_n^2 sample comes from population of two-dimensional normal dispersion (Falk 1997).

Conditions of (b)-(e) is calculated similarly to proof in (a). The positional equivalent for classical linear correlation coefficient is median correlation coefficient (Falk 1997, 1998, and Wagner 1998 et al.):

$$r_M(x,y) = CorMed(x,y) = \frac{CoMed(x,y)}{Mad(x) \cdot Mad(y)}. \qquad (8)$$

In formula (8) when median correlation is shown, only positional statistics are used, to be more precised of median for deviation observation in sample of median.

The conditions of statistics (8) are the following:

(a)r_M $(x,\ ax + b) = sign(a)$,where $sign$ (a) means function of sign for variable a,

(b)r_M $(x,\ ay + b) = sign(a) \cdot\ rM\ (x,\ y)$,

(c)$r_M(x,\ y) = 0$, where X and Y values are independent,

(d)$r_M(ax + b,\ cy + d) = sign(a) \cdot sign(b) \cdot r_M(x,\ y)$.

In case when P_n^2 sample comes from two dimensional normal dispersion with correlation coefficient ρ, then statistics r_M accepts ρ of -1 quality when $\rho = -1$, zero, when $\rho = 0$ and 1, when $\rho = 1$ (Falk 1998). It does not necessary mean that for any two-dimensional samples of r_M qualities are in range $\langle -1,\ 1\rangle$.

Example 5. Monte Carlo experiment was led concerning samples taken from one-line dispersion of general population for range (0,-1). In this case the generator of random numbers of EXCEL was used. n = 10 and 20 samples were accepted, and they were generated by M = 1000 times. The results for n = 10 samples are presented in correlative graph, where the x-axis indicated qualities of linear correlation coefficients, and the *y*-axis- median (fig.11)

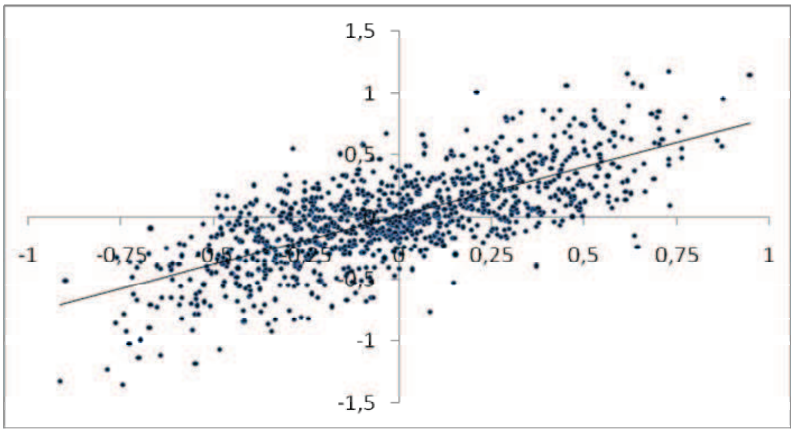

Fig.11. Correlative graph of linear and median correlation coefficients
Source: Own elaboration.

Figure 11 shows that qualities of linear correlation coefficients are in range (-1, 1), in median correlation they are out of this range and there are 8 negative and 7 positive correlation coefficients:

–1,342	–1,316	–1,225	–1,182	–1,131	–1,110	–1,065	–1,019
1,007	1,051	1,068	1,084	1,148	1,161	1,176	

By leading the simulation for n = 20 samples of one-line dispersion every median correlations are in range $\langle -1,\ 1\rangle$.

By indicating the simulation for n= *20* sample of normal dispersion of N (0, 1), EXCEL "Data Analyze" was led with normal random numbers generator' help and two correlation histograms were made (fig.12 and 13). Correlation coefficients in figures 12, 13 are included in range (-1, 1), but linear correlation range is bigger. Histogram in figure 13 shows the function of density of normal dispersion, in fig. 13 Laplace' dispersion (Fisz 1967, and Wagner 2007).

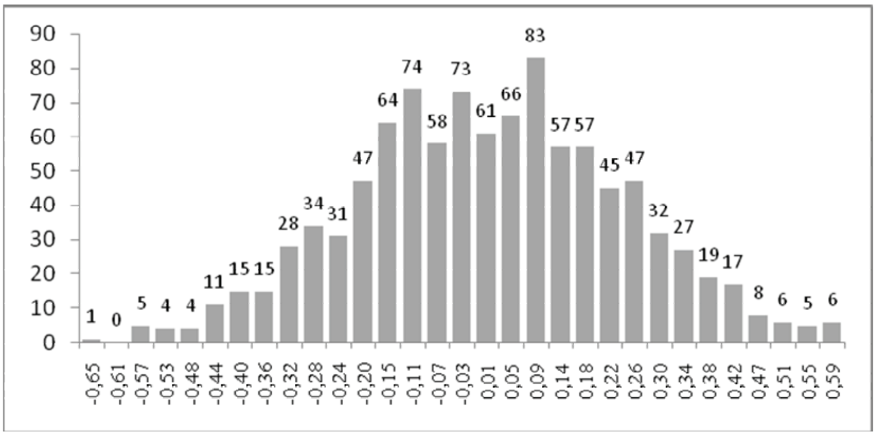

Fig. 12. Histogram of simulation of linear correlation

Source: Own elaboration.

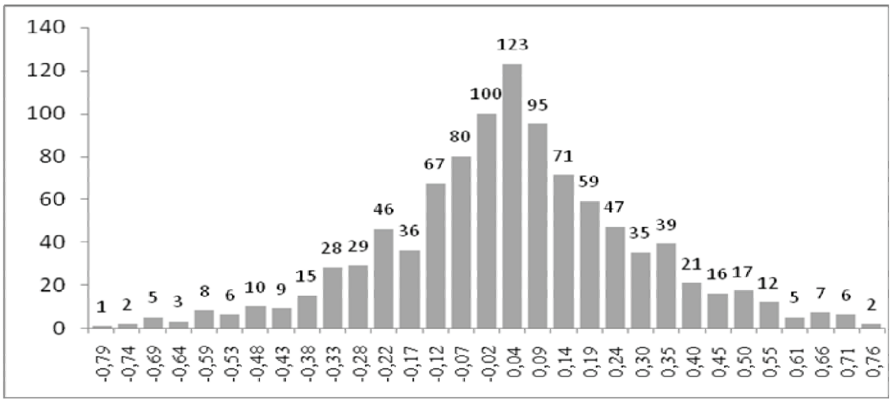

Fig. 13. Histogram of simulation of median correlation

Source: Own elaboration.

Generally for big samples and symmetric dispersion it is expected that median correlations are in range $\langle -1,\ 1\rangle$.

V. MEDIAN CORRELATION MATRIX

The equivalent of linear correlation coefficient matrix is median correlation coefficient matrix. Rm. Median correlation qualities of X_j, X_k pairs determines as follows:

a) by determining Med_j, Med_k and median absolute deviation Mad_j, Mad_k,, where Mad_j, Mad_k,, $Mad = Med(|\, x_i - Med\, |)$,

b) by establish qualityies of deviations from $X_j - Med_j,\ X_k - Med_k$,

c) by multiplying $(X_j - Med(X_j) \cdot (X_k - Med(X_k)),$

d) median is determined of given products that leads to median covariance $CoMed(X_j, X_k) = Med\{(X_j - Med_j)(X_k - Med_k)\}$,

e) matrix elements Rm are set by median covariance quotient and median absolute deviation quotient: $CorMed(X_j, X_k) = \dfrac{CovMed(X_i, X_k)}{Mad_j \cdot Mad_k}$.

Example 7. For 14 described variables, low triangular correlation coefficient median matrix is:

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}
X_1	1,000													
X_2	0,211	1,000												
X_3	0,011	0,015	1,000											
X_4	0,045	0,024	-0,022	1,000										
X_5	0,113	0,168	0,099	0,036	1,000									
X_6	0,056	0,198	0,040	0,061	0,865	1,000								
X_7	0,005	-0,102	-0,220	0,126	-0,196	-0,082	1,000							
X_8	0,758	0,309	-0,024	0,221	0,154	0,043	-0,017	1,000						
X_9	0,095	-0,134	0,000	-0,038	-0,061	-0,026	0,072	0,002	1,000					
X_{10}	-0,703	0,025	0,000	-0,073	0,068	-0,055	-0,190	-0,204	-0,392	1,000				
X_{11}	-0,014	0,138	0,045	-0,092	0,211	0,106	-0,229	0,040	-0,118	0,083	1,000			
X_{12}	-0,128	-0,039	0,091	-0,042	0,099	0,050	-0,041	-0,106	0,013	0,096	0,009	1,000		
X_{13}	0,147	0,516	-0,010	0,164	0,343	0,306	-0,036	0,246	0,018	-0,026	0,073	0,041	1,000	
X_{14}	0,190	0,500	0,000	0,197	0,315	0,242	-0,094	0,350	-0,307	0,034	0,301	-0,028	0,344	1,000

Source: Own elaboration.

The following analyses show that linear correlation coefficient matrix can be repeated. Because of their extensiveness they are not presented here. The access is limited only to comparison of both correlation coefficient matrixes. Correlation median histogram shows left- side asymmetry (Fig. 14).

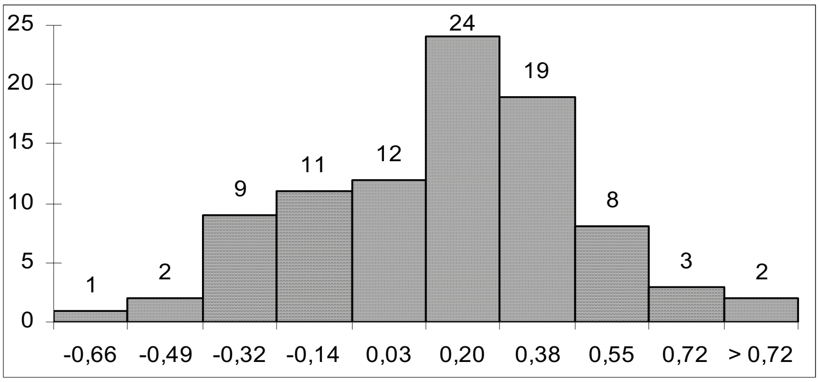

Fig.14. Histogram of median correlation coefficient number
Source: Own elaboration.

Population histogram does not differ from linear correlation coefficient population histogram both in central part as well as at the edge. Median correlations differ from linear correlations, though difference qualities are not so big, sometimes they have different signs, what is presented below.(Fig 15)

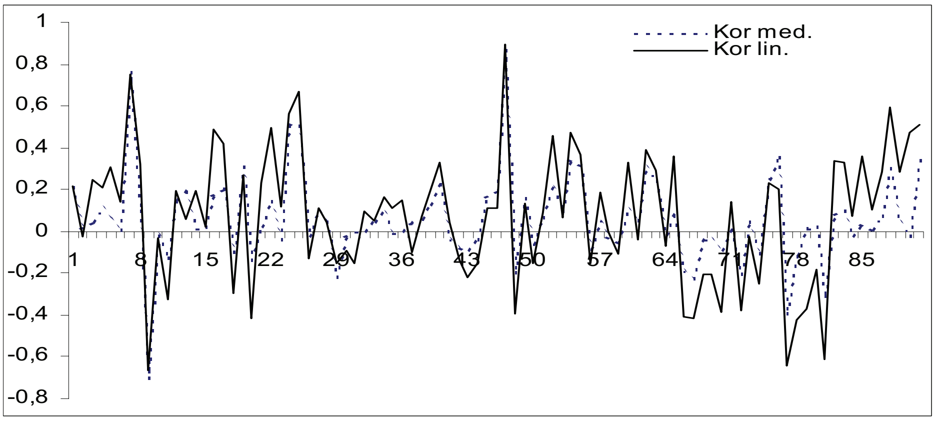

Fig.15. Comparison of linear and median correlation coefficients
Source: Own elaboration.

In general median correlation coefficients are smaller than linear correlation coefficients, and signs variance is shown in 14 cases. For analyzing deviation quantities between the mentioned coefficients, absolute difference quantities were determined, which are organized in increasing way, and the biggest 10 of them are shown below

Pair value		Correlation		Sign	Differences
		median	linear		
X_{14}	X_7	–0.0941	–0.3845	1	0.2904
X_9	X_7	0.0723	0.3628	1	0.2905
X_{14}	X_{11}	0.3010	0.5955	1	0.2945
X_{14}	X_9	–0.3074	–0.6110	1	0.3036
X_{11}	X_9	–0.1181	–0.4257	1	0.3076
X_5	X_2	0.1680	0.4905	1	0.3225
X_{14}	X_{10}	0.0343	0.3627	1	0.3284
X_{11}	X_2	0.1380	0.4946	1	0.3565
X_{12}	X_9	0.0129	–0.3734	0	0.3863
X_{13}	X_{12}	–0.0278	0.4709	0	0.4987

In column “sign” l means conformity of median and linear correlation signs, whereas 0 means, their opposite. There is a high difference in correlation coefficient quantities for (X_{12}, X_{13}), what is explained in correlative graph fig. 16. Linear correlation surface is noticed, what provokes two side diverging observations in accordance with quantity increase of both these qualities. They greatly influence average quantity, and in this way they are shown as a model for linear correlation coefficient. If these two points are skipped, there is lack in correlation between considered qualities what leads to very low median correlation quality near zero.

Similar situation appears in case of x9, x10 shown in fig. 17. Two strong matters of diverging observation appear and deform the analyzed dependency. Median correlation calculated on the basis of median is resistant to diverging observations and the influence of upper state of analyzed set, having 139 observations cause that the samples do not show correlative relation, so positive median correlation coefficient is very low.

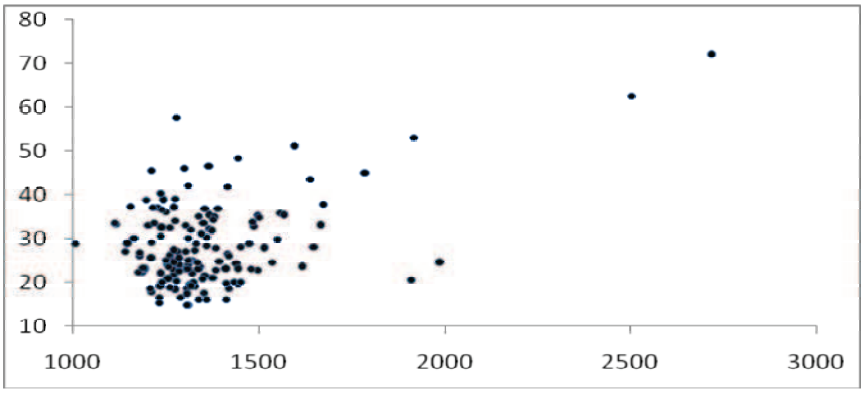

Fig . 16. Correlative graph for X_{12} i X_{13} values
Source: Own elaboration.

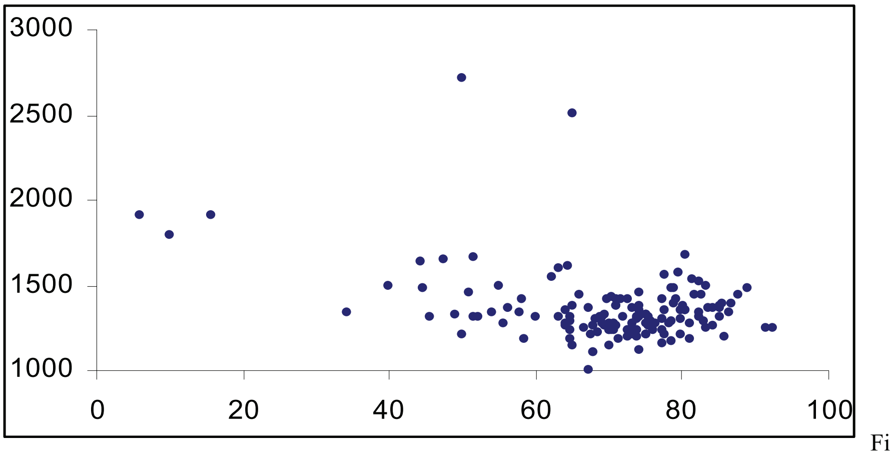

Fig.17. correlative graph for X_9, X_{10} values
Source: Own elaboration.

The examples concerning correlations between qualities present the careful usage of linear correlation coefficients. There should be preliminary research of diverging observations. Similar effects of diverging observations can be noticed when multidimensional regressive equations are estimated. The median correlations' good point is their resistance to diverging observations, because the point of dispersion of median quality is used. It is positional equivalent of classical linear correlation coefficient which calculation is based on mean, of very low point of dispersion quality.

Diagonal elements of inverse matrix and median correlative matrix are presented below:

1	2	3	4	5	6	7	8	9	10	11	12	13	14
22,991	1,971	1,079	1,507	7,610	6,378	1,299	10,382	2,197	12,683	1,256	1,053	1,622	2,032

They do not differ, though they are higher for X1, X8, X10, pointing at high dependency between these qualities

BIBLIOGRAPHY

Falk M. (1997): *On MAD and comedian.* Ann. Inst. Studia Math. 49, 615–644.

Falk M. (1998): *A Note on the Comedian for Eliptical Distribution*. Journal of Multivariate Analysis 67, 306–317.

Fisz M. (1967): *Rachunek prawdopodobieństwa i statystyka matematyczna.* PWN, Warszawa.

Mantaj A., Wagner W. (2007): *Comparative analysis of number characteristics of selected social-economic characteristics of communes of podkarpackie province.* ACTA UNIVERSITATIS LODZIENSIS, Folia Oeconomica 206, 445–461.

Rousseeuw P.J., Leroy A. (1987): *Robust Regression and Outlier Detection.* Wiley, New York.

Wagner W. (2007): *Distribution of linear combination the sample mean and the sample median.* ACTA UNIVERSITATIS LODZIENSIS, Folia Oeconomica 216, 291–302.

Wagner W., Błażczak P., Lira J. (1998): *Porównanie na materiale empirycznym charakterystyk liczbowych opartych na średniej arytmetycznej i medianie.* Listy Biometryczne XLV, 100–105.

Andrzej Mantaj, Robert Pater, Wiesław Wagner

ASPEKTY INTERPRETACYJNE MACIERZY WSPÓŁCZYNNIKÓW KORELACJI LINIOWEJ I MEDIANOWEJ

Macierz współczynników korelacji liniowej odgrywa podstawową rolę w badaniu zależności układu cech w wielowymiarowej analizie statystycznej. Pozwala ona na określenie cech mocno skorelowanych w oparciu o wartości diagonalne jej macierzy odwrotnej, a także na interpretację zmiennych przy wykorzystaniu statystyk liczbowych i wykresów zbudowanych na jej elementach.

Alternatywą wspomnianej macierzy jest macierz korelacji medianowej. Jest ona propozycją, w przypadku gdy niektóre z nich wykazują występowanie obserwacji odstających, co często ma miejsce w analizie zjawisk społeczno-gospodarczych. Powstają wówczas trudności z właściwą oceną powiązań cech. Jest to powodowane niską wartością załamania średniej arytmetycznej. Dlatego w miejsce średniej arytmetycznej jako klasycznej charakterystyki liczbowej położenia, stosuje się pozycyjną charakterystykę liczbową – medianę, która z kolei ma bardzo wysoką wartością załamania.

Celem pracy jest porównanie współczynników korelacji liniowej i medianowej na materiale empirycznym. Podano różne wzory na wyznaczanie współczynnika korelacji liniowej oraz określono podstawowe własności medianowego odchylenia bezwzględnego w przypadku próby jednowymiarowej i kormedian dla próby dwuwymiarowej. Te dwie ostatnie charakterystyki wykorzystuje się przy obliczaniu współczynnika korelacji medianowej, którego własności także zaprezentowano w pracy.

Materiał badawczy stanowią dane liczbowe 14 cech społeczno-gospodarczych dla gmin wiejskich i miejsko-wiejskich województwa podkarpackiego według stanu na dzień 31.12.2002 r.