APPLICATION OF CORRELATION COEFFICIENTS IN QUALITATIVE DATA ANALYSIS

Krzysztof Buczkowski,
Warsaw University of Technology, Department of Cartography, Poland, Warszawa, Pl. Politechniki 1, 00-668, Instytut Fotogrametrii i Kartografii PW
k.buczkowski@gik.pw.edu.pl tel: +(48) 22 660-73-09; fax: +(48) 22 629-91-82

Robert Olszewski,
Warsaw University of Technology, Department of Cartography, Poland, Warszawa, Pl. Politechniki 1, 00-668, Instytut Fotogrametrii i Kartografii PW
r.olszewski@gik.pw.edu.pl tel: +(48) 22 660-73-09; fax: +(48) 22 629-91-82

The information processing functions available in geographic information systems allow only to make simple spatial analyses such as: layer crossing, object selection by a defined attribute or within a created equidistant, etc. Therefore, the aim of cartography becomes searching for new methods allowing to optimally use all information in the databases. The method [K.A. Saliszczew 1955, A.M. Berlant 1986] of significant importance here is the cartographic analysis. It allows to create and use study procedures to analyse spatial phenomena and to present the results in a cartographic form. One of the important aims of the method is analysing correlation between phenomena. There are many ways of carrying such studies, depending on specification of source data, statistic measures and the type of procedures. In the present work only some aspects of that complex and extraordinary problem are discussed. The authors investigate the possibility of applying Yule's qualitative attributes correlation coefficient, also called tetrachoric correlation coefficient, into analysis of correlation between two phenomena of qualitative character.

Attributes correlation coefficient is based on a popular non-parametric test χ^2. It is used in testing hypotheses [G.B. Norcliffe 1986], analysing distribution normality [C. Domański 1990] and comparing two or more structures [Z. Barańska 1995]. It is also the basis for calculating a correlation coefficient allowing to define the degree of correlation between two qualitative phenomena and to evaluate the importance of the correlation. One can find some information about cartographic aspects of using qualitative attributes correlation coefficient in the works of [A.M. Berlant 1978].

The basics of chi-square test χ^2

Theoretical basis of χ^2 test is very obvious. Let's consider that the empirical distribution of a set with *m* elements is known (Table 1).

Table 1. The empirical distribution for a set with *m* elements

Values	Empirical frequencies
x_1	n_1
x_2	n_2
.....	.....
x_m	n_m

Let's assume then that we can characterise the theoretical distribution of the considered empirical distribution (Table 2).

Table 2. The theoretical distribution for the empirical distribution of Table 1

Empirical frequencies	Probability	Theoretical frequencies
n_1	p_1	$p_1 \cdot n$
n_2	p_2	$p_2 \cdot n$
.....	.....	.....
n_m	p_m	$p_m \cdot n$
$\sum_{i=1}^{m} n_i = n$	$\sum_{i=1}^{m} p_i = 1$	$\sum_{i=1}^{m} p_i \cdot n = n$

Let's compare empirical and theoretical frequencies. The simplest way for it is to describe differences between them:

$$n_1 - p_1 \cdot n$$
$$n_2 - p_2 \cdot n$$
$$.....$$
$$n_m - p_m \cdot n.$$

The sum of counted differences equals zero. In order to avoid zero as the result it is necessary to take as a measure of difference between frequencies the following expression:

$$\sum_{i=1}^{m} (n_i - p_i \cdot n)^2. \qquad (1)$$

The expression (1) must be standardised. It is connected with an obvious fact that it is not possible to give the same weights to differences like these given in the following example: $(5\text{-}3)^2$ and $(100 - 98)^2$. Let's make standardisation dividing one by one individual elements of a series in order by the proper value of a theoretical frequency. Then we will obtain the expression:

$$\sum_{i=1}^{m} \frac{(n_i - p_i \cdot n)^2}{p_i \cdot n} \; . \qquad (2)$$

At $n \longrightarrow \infty$ the expression (2) is convergent to χ^2 distribution with (m-1) degrees of freedom, where m is a number of classes, under the condition that frequencies of individual classes are bigger or equal 10. The χ^2 statistics assumes values with the interval:

$$\left[0, n \cdot \sqrt{(m-1)}\right].$$

If $\chi^2 = 0$, empirical frequencies are equal to theoretical ones. The bigger is the χ^2 value the more important is the difference between empirical and theoretical frequencies.

Let's consider the simplest case of two dichotomic phenomena given conventionally as *X, Y*. Let's suppose that the phenomenon *X* has the feature x_1 or x_2, and the phenomenon *Y* has the feature y_1 or y_2. For simplification the considered empirical distribution can be written in so called correlation table (Table 3).

Table 3. The empirical distribution of the *X, Y* phenomena

Y \ X	x_1	x_2	total
y_1	*a*	*b*	n_1
y_2	*c*	*d*	n_2
total	n_3	n_4	*n*

x_1, x_2 - features concerning the *X*
y_1, y_2 - features concerning the *Y*
a, b, c, d - empirical frequencies

For the empirical distribution introduced in Table 3 can be given the theoretical distribution, shown in Table 4.

Table 4. The theoretical distribution of phenomena *X,Y* occurrence

Y \ X	x_1	x_2	total
y_1	$(n_3 \cdot n_1)/n$	$(n_4 \cdot n_1)/n$	n_1
y_2	$(n_3 \cdot n_2)/n$	$(n_4 \cdot n_2)/n$	n_2
total	n_3	n_4	n

x_1, x_2 - features connected with the X
y_1, y_2 - features connected with the Y

On the basis of the empirical and theoretical distributions, according the expression (2), we can determine the χ^2 value.
Because:

$$\chi^2 = \frac{\left(a - \frac{n_3 \cdot n_1}{n}\right)^2}{\frac{n_3 \cdot n_1}{n}} + \frac{\left(b - \frac{n_4 \cdot n_1}{n}\right)^2}{\frac{n_4 \cdot n_1}{n}} + \frac{\left(c - \frac{n_3 \cdot n_2}{n}\right)^2}{\frac{n_3 \cdot n_2}{n}} + \frac{\left(d - \frac{n_4 - n_2}{n}\right)^2}{\frac{n_4 \cdot n_2}{n}} \qquad (3)$$

and $\left|a - \frac{n_3 \cdot n_1}{n}\right| = \left|b - \frac{n_4 \cdot n_1}{n}\right| = \left|c - \frac{n_3 \cdot n_2}{n}\right| = \left|d - \frac{n_4 \cdot n_2}{n}\right| = \left|\frac{a \cdot d - b \cdot c}{n}\right|$

the χ^2 value after obvious transformations can be written in the form:

$$\chi^2 = \frac{n \cdot (a \cdot d - b \cdot c)^2}{n_1 \cdot n_2 \cdot n_3 \cdot n_4}, \qquad (4)$$

where: $n_1 = a + b$, $n_2 = c + d$, $n_3 = a + c$, $n_4 = b + d$.

The χ^2 value has been determined for the correlation table consisting of two rows and two columns. In the case of examining correlation concerning bigger number of phenomena features the table containing bigger number of rows and columns and the expression for it can be written in the similar way. In the general case when a table consists of k rows and l columns, the χ^2 statistics has values from the interval:

$$\left[0, n \cdot \sqrt{(k-1) \cdot (l-1)}\right].$$

Yule's correlation coefficient

The χ^2 statistics is the basis for counting the quality features correlation coefficient. For two dichotomic phenomena, the χ^2 has values from the interval [*0, n*]. Dividing χ^2 by n normalises the range of values, reducing their variation to the interval [*0, 1*]. Determining square root of the χ^2/n expression, gives the change of the values range to the interval [*-1, 1*]. In this way we obtain the formula for feature correlation coefficient for two dichotomic phenomena, so called Yule's correlation coefficient:

$$\phi = \sqrt{\frac{\chi^2}{n}}. \qquad (5)$$

After introducing the determined χ^2 value we obtain the following form of this formula useful for practical use:

$$\phi = \frac{a \cdot d - b \cdot c}{\sqrt{(a+b) \cdot (a+c) \cdot (c+d) \cdot (b+d)}}. \qquad (6)$$

In the case independence of phenomena features: $\varphi = 0$ or equals almost zero; and in the case of functional correlation of features: $\varphi = \pm 1$. All intermediate states give information about positive or negative level of dependence for the phenomena features.

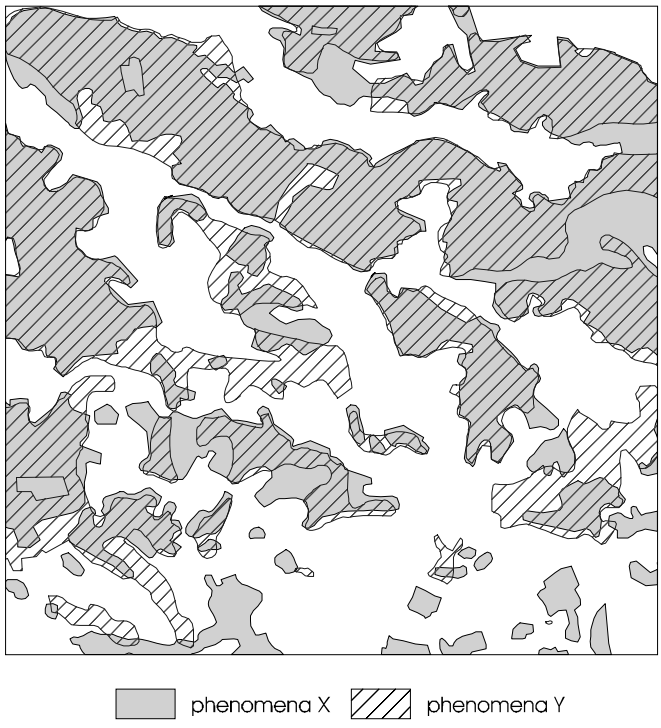
From the formula (5) is known that the value of coefficient φ can be obtained directly from the χ^2 value and vice versa. These two measures have distinct character and each of them gives different information. The correlation coefficient φ informs about strength of connection between features of phenomena while the χ^2 test informs about importance of this connection. One can conclude that this connection can be important but weak or strong but irrelevant.

Application of Yule's correlation coefficient in the method of cartographic analysis

As mentioned above, A.M. Berlant [1978] was the first one who tried to apply qualitative attributes correlation coefficient in the analysis of phenomena of spatial character. He discussed the issue of qualitative attributes correlation (tetrachoric correlation) as a specific case of polichronic correlation allowing to evaluate the degree of correlation between two phenomena measured in a rang scale. However, he noticed significant drawbacks of a correlation coefficient. In this approach coefficient value mainly depends on the size of area in which none of phenomena is present that is on the size of a source map sheet and its edition – content to background ratio.

The approach suggested by the authors of the present paper allows to eliminate the drawbacks and to define spatial diversification of degree of correlation between phenomena [K. Buczkowski, R. Olszewski 2000].

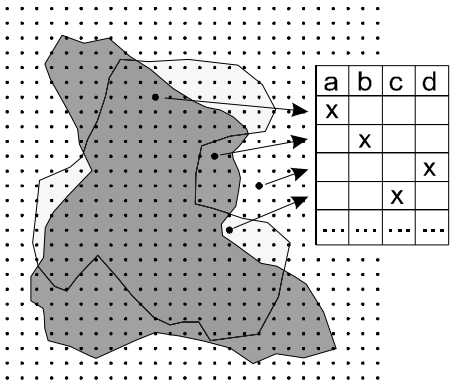
Let's assume that in some area two dychotomic phenomena , X and Y measured in qualitative scale (picture 1), are present in an island form. As the analysis is of a methodical character it is more convenient to assume a model distribution of phenomena in order to consider all possible cases.



Picture 1. Distribution of analysed phenomena *X* and *Y*

For practical reasons, during calculation of correlation coefficient it is easier to partially replace continuous way of phenomena occurrence with a discrete model. The best solution is to cover

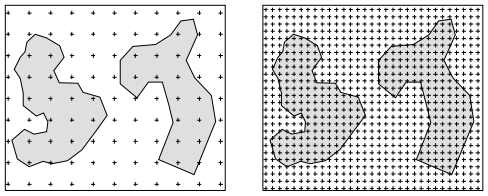
the map with a net of regularly distributed points so that they precisely represent the analysed phenomena (picture 2).



Picture 2. The discrete model and description attributes

To each point of the model one can ascribe one of the four states:
a – both phenomenon X and phenomenon Y occur,
b –only phenomenon X occurs,
c - only phenomenon Y occurs,
d – none of phenomena occurs.
If we then ascribe frequency to each of the states, we will define the empirical distribution allowing to calculate a correlation coefficient ϕ.
In the analysed case the dychotomic division occurs and that is why Yule's correlation coefficient is expressed by the formula (6).

Density of the net points is an extremely significant issue as it influences both the coefficient value and the probability of its occurrence (picture 3).



Picture 3. Replacement of partially contiguous phenomena occurrence with a discrete model.
a – insufficiently precise model (2601 points for the whole set from picture 1)
b – sufficiently precise model (18000 points for the whole set from picture 1)

In the table 5 there are presented results of calculation of correlation coefficient value and the related probability for different density of net points. The calculation is done for all net points therefore for the whole area of the map.

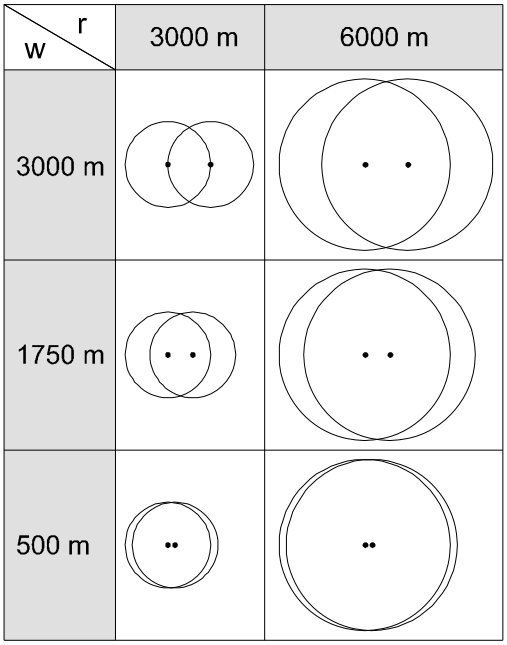
Table 5. Correlation coefficient and probability of correlation occurrence in the whole set for different number of net points

number of net points	correlation coefficient ϕ	probability of correlation occurrence
121	*0,243*	*0,340*
729	*0,388*	*0,447*
2601	*0,564*	*0,941*
10609	*0,598*	*0,998*
15625	*0,590*	*0,999*

As expected, the increasing number of points results in the increase of the confidence level towards calculated correlation measures. For enough big number of points, for example *10609,* we can assume with certainty of *99,8%* that the correlation coefficient value is *0,598*. Starting with *15000* points the coefficient begins to have the constant value of *0,590*, identical with the result of calculation based on area covered by phenomena. Therefore, one can assume that the density of the discrete model consisting of such a number of points is sufficient to represent the analysed phenomena. In the further discussion a model of a bigger density, consisting of 18000 points, is analysed. Such a density of the model assures sufficient reliability of the analysis.
The model represents partially continuous phenomena distribution with a satisfactory precision. The attributes are ascribed to each point according to the rules presented in picture 2.

Applying a circle base area in the analysis

In order to define correlation between phenomena in different space locations circle base areas of different size and translation vectors are used (picture 4).



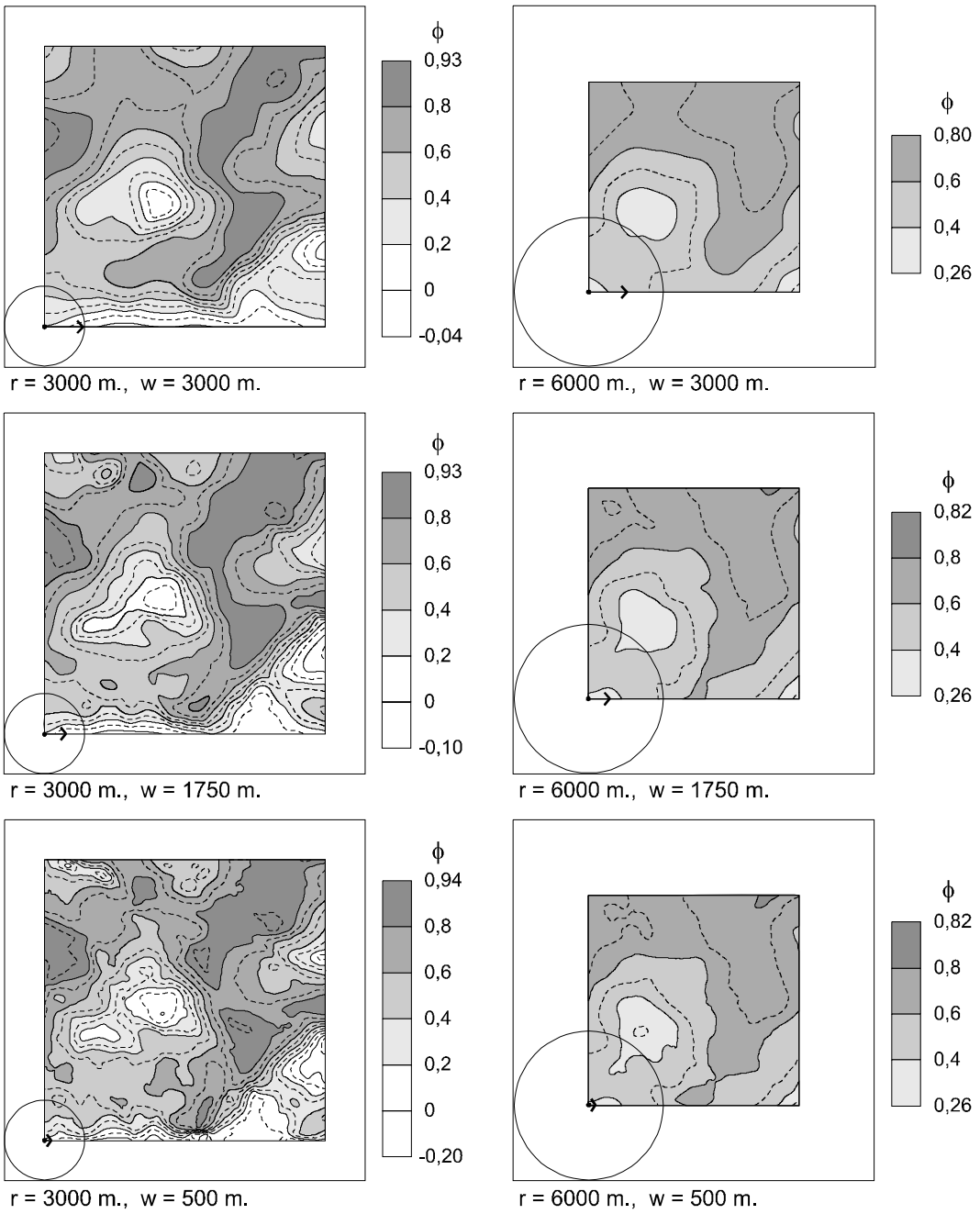
Picture 4. Sizes and translation vectors of circle base areas.

Using a circle base area while creation of isoline maps is to some extent controversial. W.A. Czerwiakow [1975, 1978] believes that a circle base area being the most homogeneous figure is the best basis to measure the value of phenomenon occurring around any chosen point (circle centre). J.

Mościbroda [1999] thinks that "each separate measure done for different locations of the circles seems from theoretical point of view, the most reliable but their unification in order to create a map creates mistakes".

Applying a circle base area in the analysis of correlation in different space locations seems to be the best solution. Such an area because of its shape optimises the choice of model points. The coefficient value calculation is based on a constant number of points according to the formula (6), not on making average of the values. The calculated value is ascribed to the circle centre which is a definite space point. By changing the size of the translation vector one can increase or decrease the number of space points for which the coefficient value is calculated.

Let's analyse now how the results change depending on the changes in the size of the area and translation vector (picture 5).



Picture 5. Correlation between occurrence of phenomena X i Y in cases of different size of

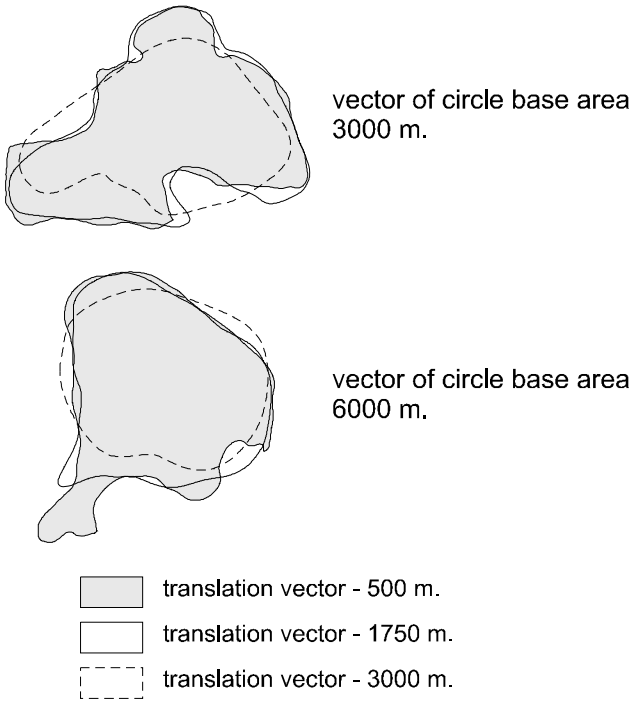
circle base area and translation vector

The analysis is done using circle base areas of two radius sizes: 3 and 6 km.
In the smaller area 708 model points are included while in the bigger one 2832 points are included. The number of points in the basic areas allows to achieve a very big reliability of study in both cases.
The increase in the size of a basic area obviously results in generalisation of the achieved results. The generalisation concerns not only isoline course, but also radically lessens the range of correlation coefficient value for the whole set (table 6).

Table 6. Influence of size and vector of circle base area on the range of correlation coefficient value

Circle area radius	Translation vector	Minimal value φ	Maximal value φ
6000 m	500 m	0,26	0,82
6000 m	1750 m	0,26	0,82
6000 m	3000 m	0,26	0,80
3000 m	500 m	-0,20	0,94
3000 m	1750 m	-0,10	0,93
3000 m	3000 m	-0,04	0,93

When the area size is defined the change in the length of translation vector does not significantly influence the range of correlation coefficient value, but it changes generalisation of isoline course. When translation is small, for instance 500 m., isoline course is complex and it allows for small, local differences in the coefficient value (picture 6).



Picture 6. Isoline course 0,2 for different size of basic area and different length of translation vector

The increase of the translation vector results in generalisation of isoline course. It becomes smoother and smoother and it shows only general tendencies of spatial changes in coefficient value

(picture 6). During maps preparation it is assumed that the length of the biggest translation vector equals the radius of the smaller circle area and it is 3000 m. It is also assumed that the smallest vector is 500 m. Further decrease in the length of the vector does not cause any significant changes in the isoline course.

Summary

- In the present paper it was proved that the qualitative attributes correlation coefficient can be used to define correlation between phenomena of spatial character.
- In the study only the simplest case was analysed that is correlation between two phenomena. One should also generalise and modify the coefficient in order to define correlation among the bigger number of phenomena.
- The analysis cannot be fully automated. It requires a conscious choice of parameters (the number of model points and area size) and right interpretation of the results
- The study of correlation between phenomena occurrence can be based not only on geometric areas but also on irregular areas such as administrative division units, drainage area, etc.
- Applying a circle base area allows to increase the number of points, for which correlation coefficient value is calculated. It creates possibility of spreading much more reliable statistical area on the analysed one resulting in a more precise and reliable isoline course.
- An extremely important issue is the right choice of basic area size. It is the main generalisation factor. It should be chosen empirically, depending on the aims of a study.

Bibliography

[1] Barańska Z.: *Podstawy metod statystycznych dla psychologów*. Gdańsk 1995. Wydawnictwo Uniwersytetu Gdańskiego.
[2] Berlant A.M.: *Kartograficzeskij mietod isslledowanija.* Moskwa 1978, Izdat. Moskowskowo Uniwersiteta.
[3] Berlant A.M.: *Obraz prastranstwa: karta i informacyja*. Moskwa 1986, Izdat. Mysl.
[4] Buczkowski K., Olszewski R.: *Zastosowanie współczynnika korelacji Yule`a do badania zależności między występowaniem zjawisk jakościowych*. Warszawa 2000. Polski Przegl. Kartogr., T.32, 2000, nr 1.
[5] Czerwiakow W.A.: *Tocznost i detalnost izoliniejnych polej plotnosti.* Moskwa 1975, Izw. Wys. Uczebn. Zaw. ,Gieod. i Aerofotosj., nr 6
[6] Czerwiakow W.A.: *Koncepcija pola w sowremiennoj kartografii.* Nowosybirsk 1978, Izd. „Nauka".
[7] Domański C.: *Testy statystyczne.* Warszawa 1990. PWE.
[8] Mościbroda J.: *Mapy statystyczne jako nośniki informacji ilościowej.* Lublin 1999. Wydawnictwo UMCS.
[9] Norcliffe G.B.: *Statystyka dla geografów*. Warszawa 1986. PWN.
[10] Saliszczew K.A.: *O kartograficzeskom mietodie issledowanija*. Wiestnik Mosk. Uniwers. Ser.5 Gieogr. 1955, nr 10.