COMPARISON OF VALUES OF PEARSON'S AND SPEARMAN'S CORRELATION COEFFICIENTS ON THE SAME SETS OF DATA

JAN HAUKE, TOMASZ KOSSOWSKI

Adam Mickiewicz University, Institute of Socio-Economic Geography and Spatial Management, Poznań, Poland

Manuscript received April 19, 2011
Revised version May 18, 2011

HAUKE J., KOSSOWSKI T., Comparison of values of Pearson's and Spearman's correlation coefficient on the same sets of data. *Quaestiones Geographicae* 30(2), Bogucki Wydawnictwo Naukowe, Poznań 2011, pp. 87–93, 3 figs, 1 table. DOI 10.2478/v10117-011-0021-1, ISBN 978-83-62662-62-3, ISSN 0137-477X.

ABSTRACT: Spearman's rank correlation coefficient is a nonparametric (distribution-free) rank statistic proposed by Charles Spearman as a measure of the strength of an association between two variables. It is a measure of a monotone association that is used when the distribution of data makes Pearson's correlation coefficient undesirable or misleading. Spearman's coefficient is not a measure of the linear relationship between two variables, as some "statisticians" declare. It assesses how well an arbitrary monotonic function can describe a relationship between two variables, without making any assumptions about the frequency distribution of the variables. Unlike Pearson's product-moment correlation coefficient, it does not require the assumption that the relationship between the variables is linear, nor does it require the variables to be measured on interval scales; it can be used for variables measured at the ordinal level. The idea of the paper is to compare the values of Pearson's product-moment correlation coefficient and Spearman's rank correlation coefficient as well as their statistical significance for different sets of data (original - for Pearson's coefficient, and ranked data for Spearman's coefficient) describing regional indices of socio-economic development.

KEY WORDS: Pearson's correlation coefficient, Spearman's rank correlation coefficient, Kendall's tau, regional indices of socio-economic development

Jan Hauke, Tomasz Kossowski, Institute of Socio-Economic Geography and Spatial Management, Adam Mickiewicz University, ul. Dzięgielowa 27, 61-680 Poznań, Poland; e-mail: jhauke@amu.edu.pl, tkoss@amu.edu.pl

1. Introduction: Historical background

Correlations between variables can be measured with the use of different indices (coefficients). The three most popular are: Pearson's coefficient (**r**), Spearman's rho coefficient ($\mathbf{r_s}$), and Kendall's tau coefficient ($\boldsymbol{\tau}$). Kendall's tau, introduced by Kendall (1938), is a correlation coefficient that can be used as an alternative to Spearman's rho for data in the form of ranks. It is a simple function of the minimum number of neighbour swaps needed to produce one ordering from another. Its properties were also analysed by Kendall in his book concerning rank correlation methods, first published in 1948. As he states there, "The coefficient we have introduced provides a kind of average measure of the agreement between pairs of members ("agreement", that is to say, in respect of order) and thus has evident recommendation as a measure of the concordance between two

rankings" (p. 7), and "In general, rho is an easier coefficient to calculate than τ. We shall see … that from most theoretical points of view τ is preferable to rho…" (p. 11).

The main advantages of using Kendall's tau are the fact that its distribution has slightly better statistical properties, and that there is a direct interpretation of this statistics in terms of probabilities of observing concordant and discordant pairs. Nonetheless, coefficient τ has not been used so often in the past (the last sixty years) as Spearman's coefficient in measuring rank correlation, mainly because it was the one more difficult to compute. Nowadays the calculation of Kendall's τ poses no problem. Kendall's τ is equivalent to Spearman's r_s in terms of the underlying assumptions, but they are not identical in magnitude, since their underlying logic and computational formulae are quite different. The relationship between the two measures for large numbers of pairs is given by Daniels (1944):

$$-1 \leq 3\tau - 2r_s \leq 1$$

In most cases, these values are very close and would invariably lead to the same conclusions, but when discrepancies occur, it is probably safer to interpret the lower value. More importantly, Kendall's τ and Spearman's r_s imply different interpretations. Spearman's r_s is considered the regular Pearson's correlation coefficient in terms of the proportion of variability accounted for, whereas Kendall's τ represents a probability, i.e., the difference between the probability that the observed data are in the same order versus the probability that the observed data are not in the same order. Properties and comparisons of Kendall's τ and Spearman's r_s have been analysed by many researchers and they are still under investigation (see e.g. Valz & Thompson 1994, Xu et al. 2010). Bearing in mind the comments mentioned above, we will treat Spearman's coefficient as the proper representative measure for ranks correlation.

The idea of the paper is to compare the values of Pearson's correlation coefficient treating data in a quantitative way versus the values of Spearman's rank correlation coefficient treating the same data in a somewhat 'qualitative' way for real data sets.

Coming back to the history of developing the idea of measuring correlation strength, one should mention the set of papers by Galton, Yule and Pearson (listed in the references) which created the basis for a proper and correct application and interpretation of correlations (in the modern meaning of the word). The history (till 1985) was presented by Rodgers & Nicewander (1988) in Table 1.

The history and properties of Pearson's correlation coefficient were also described by Pearson (1920), Weida (1927), Walker (1928), Stigler (1988), and Piovani (2008). It is worth noting that some authors use the term "Fisher's correlation coefficient", e.g. Plata (2006), as R.A. Fisher also worked in the area of correlation (Fisher 1915, 1921). His contribution was described by Anderson (1996), who mentioned another statistician interested in properties of the coefficient of correlation, namely W.S. Gosset, known as 'Student' (1908).

Pearson's coefficient of correlation was discovered by Bravais in 1846, but Karl Pearson was the first to describe, in 1896, the standard method of its calculation and show it to be the best one possible. Pearson also offered some comments about an extension of the idea made by Galton (who applied it to anthropometric data). He called this method the "product-moments'' method (or the Galton function for the coefficient of correlation r). An important assumption in Pearson's 1896 contribution is the normality of the variables analysed, which could be true only for quantitative variables. Pearson's correlation coefficient is a measure of the strength of the linear relationship between two such variables.

In 1904 Spearman adopted Pearson's correlation coefficient as a measure of the strength of the relationship between two variables that cannot be measured quantitatively. He noted: "The most fundamental requisite is to be able to measure our observed correspondence by a plain numerical symbol. There is no reason whatever to be satisfied either with vague generalities such as "large", "medium", "small," or, on the other hand, with complicated tables and compilations. The first person to see the possibility of this immense advance seems to have been Galton, who, in 1886, writes: …" (Spearman 1904a: 73–74).

Table 1. Landmarks In the History of Correlation and Regression.

Date	Person	Event
1823	Carl Friedrich Gauss, German mathematician	Developed the normal surface of N correlated variates.
1843	John Stuart Mill, British philosopher	Proposed four canons of induction, including concomitant variation.
1846	Augusts Bravais. French naval officer and astronomer	Referred to "une correlation", worked on bivariate normal distribution.
1868	Charles Darwin, Galton's cousin, British natural philosopher	"All parts of the organisation are ... connected or correlated."
1877	Sir Francis Gallon, British, the first biometrician	First discussed "reversion", the predecessor of regression.
1885	Sir Francis Gallon	First referred to "regression". Published bivariate scatterplot with normal isodensity lines, the first graph of correlation. "Completed the theory of bi-variate normal correlation." (Pearson 1920)
1888	Sir Francis Gallon	Defined *r* conceptually, specified its upper bound.
1895	Karl Pearson, British statistician	Defined the (Galton-) Pearson product-moment correlation coefficient.
1920	Kart Pearson	Wrote "Notes on the History of Correlation".
1985		Centennial of regression and correlation.

Spearman's rank correlation coefficient is a nonparametric (distribution-free) rank statistic proposed as a measure of the strength of the association between two variables. It is a measure of a monotone association that is used when the distribution of data makes Pearson's correlation coefficient undesirable or misleading. Spearman's coefficient is not a measure of the linear relationship between two variables, as some "statisticians" declare. It assesses how well an arbitrary monotonic function can describe the relationship between two variables, without making any assumptions about the frequency distribution of the variables. Unlike Pearson's product-moment correlation coefficient, it does not require the assumption that the relationship between the variables is linear, nor does it require the variables to be measured on interval scales; it can be used for variables measured at the ordinal level. In principle, $\mathbf{r}_s$ is simply a special case of Pearson's product-moment coefficient in which the data are converted to ranks before calculating the coefficient. It should be noted that Spearman made an error in his correlation formula on page 77: he used medians instead of means in the definition of $\mathbf{r}_s$. He corrected this in his later works. Spearman's statistical accomplishments of 1904 were not appreciated by his University College colleague Karl Pearson, and there was a long-standing disagreement between them. The history and subsequent practice showed that it was Spearman who was right, and nowadays coefficient $\mathbf{r}_s$ is widely used in statistical analyses.

The use of Pearson's product-moment correlation coefficient and Spearman's rank correlation coefficient for geographical data (on map data that are spatially correlated) was examined by Haining (1991); see also Griffith (2003).

In the present paper we would like to compare the values and significance of Pearson's and Spearman's coefficients for the same sets of data (original data for $\mathbf{r}$ and ranked data for $\mathbf{r}_s$). The data used in the analysis represent regional indices of socio-economic development.

2. Calculation results

Basis for calculation: 1998 data of the Central Statistical Office for chosen administrative units of different levels in Poland, with the use of the following variables:

- X1 Population by official place of residence
- X2 Telephones per 1000 population
- X3 Water supply: amount of water supplied to households
- X4 Density of population per 1 sq. km
- X5 Arable land by administrative borders
- X6 Commune area in sq. km

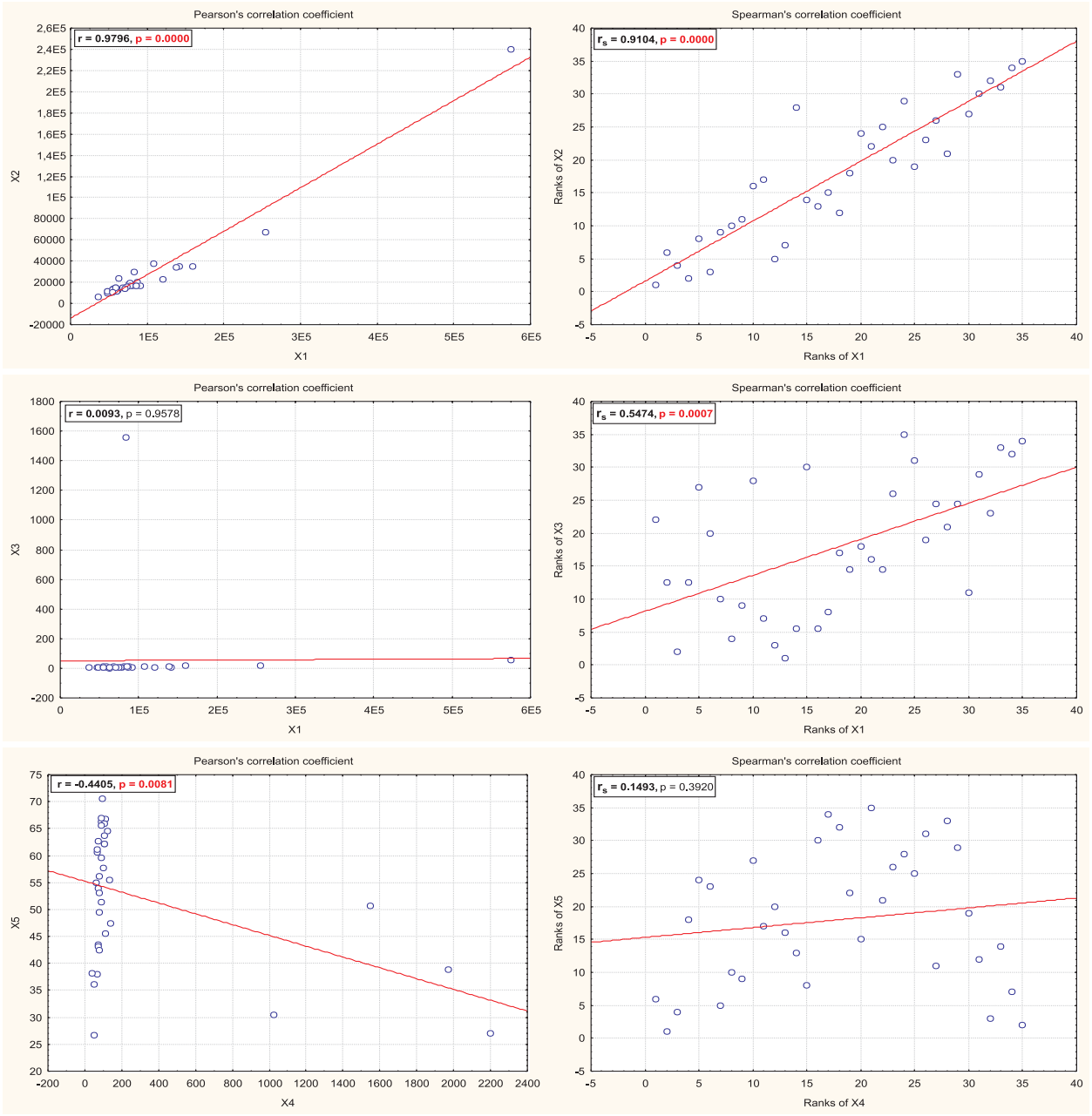

Fig. 1. Comparison of Pearson's and Spearman's coefficients of correlation of selected variables for poviats of Wielkopolska voivodeship.

- X7 Employable population of working age (18–64 for males, 18–59 for females)
- X8 Permanent migration rate per 1000 population
- X9 Industrial employment per 1000 workers
- X10 Live births per 1000 population
- X11 Consumption of water in national economy
- X12 Birth rate, total

We used these data to calculate Pearson's and Spearman's correlation coefficients. The analysis was divided into three parts, depending on the spatial scale of the variables. At the first level of analysis we used n=35 subregions (poviats) in Wielkopolska voivodeship. In studying this area, we calculated three pairs of correlation coefficients for the following variables: X1–X2, X1–X3, and X4–X5. Observe (Fig. 1) that for the first pair both Pearson's and Spearman's correlation coefficients are high and highly significant. In the case of the second pair, only Spearman's coefficient is significant, and in the third case we only

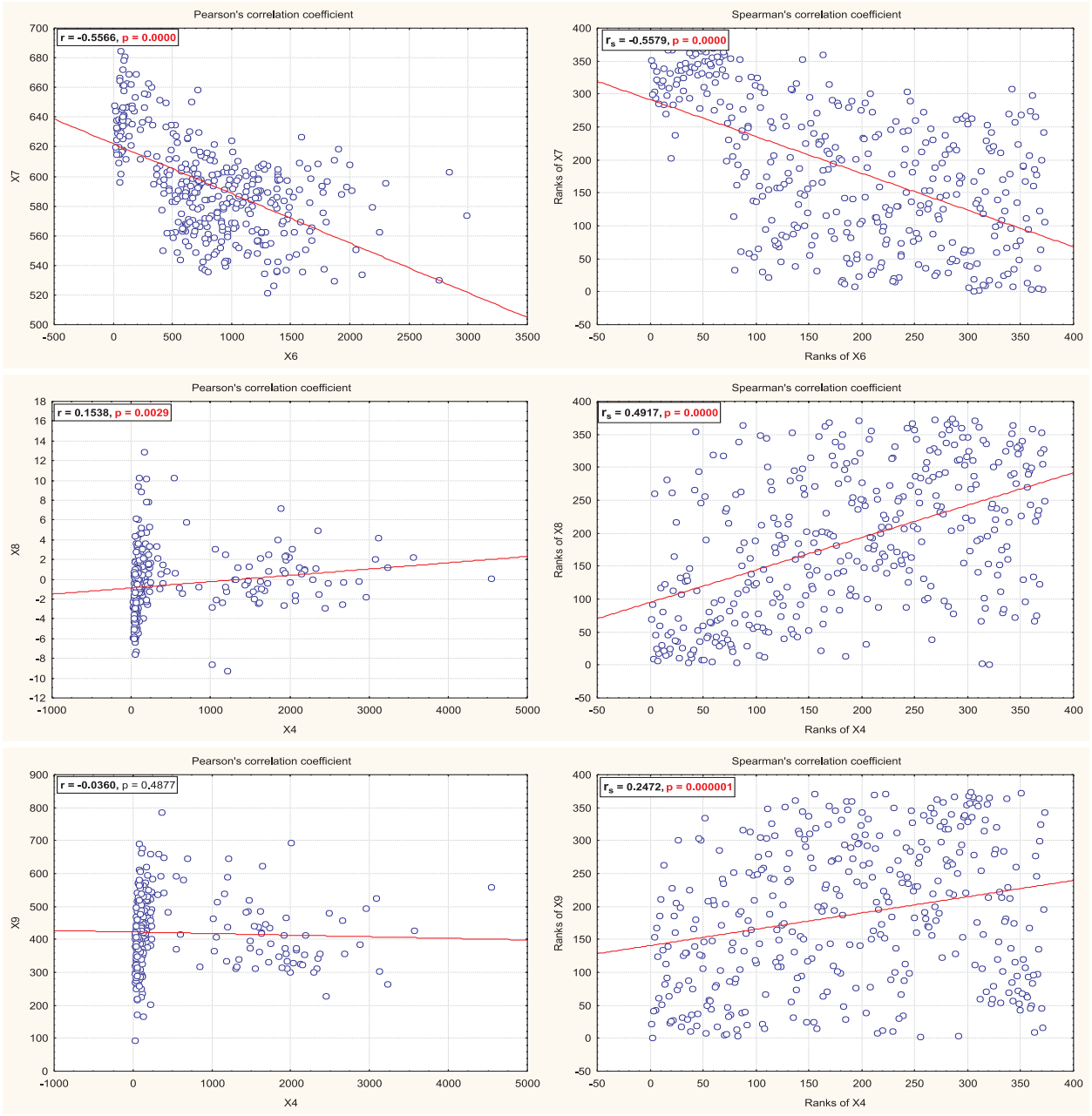

Fig. 2. Comparison of Pearson's and Spearman's coefficients of correlation of selected variables for poviats in Poland.

have significance for Pearson's coefficient. It is interesting to note that in the last case we have two different trends in the data, but only one of them is significant.

The second group of pairs (the second level) was obtained for the subregional level again, but in the whole of Poland, with n=373 (Fig. 2). Both coefficients were highly significant for the first pair X6–X7 and equalled about –0.56. In the case of the second pair we found two significant correlations between X4 and X8, but Spearman's coefficient was higher than Pearson's. The last pair in this series was X4–X9. In this case we have got Pearson's coefficient insignificant and negative but close to zero, while Spearman's coefficient was significant and equal to 0.25.

Figure 3 presents the result for the commune level in Poland (the third level of analysis) where the number of units is n=3,056 (we divided mixed urban-rural units into the urban and the rural part). For the first (X1–X10) and the second (X1–X11) pairs, we obtained positive and highly significant values of Pearson's and Spearman's correlation coefficients. For the last pair (X12–X7)

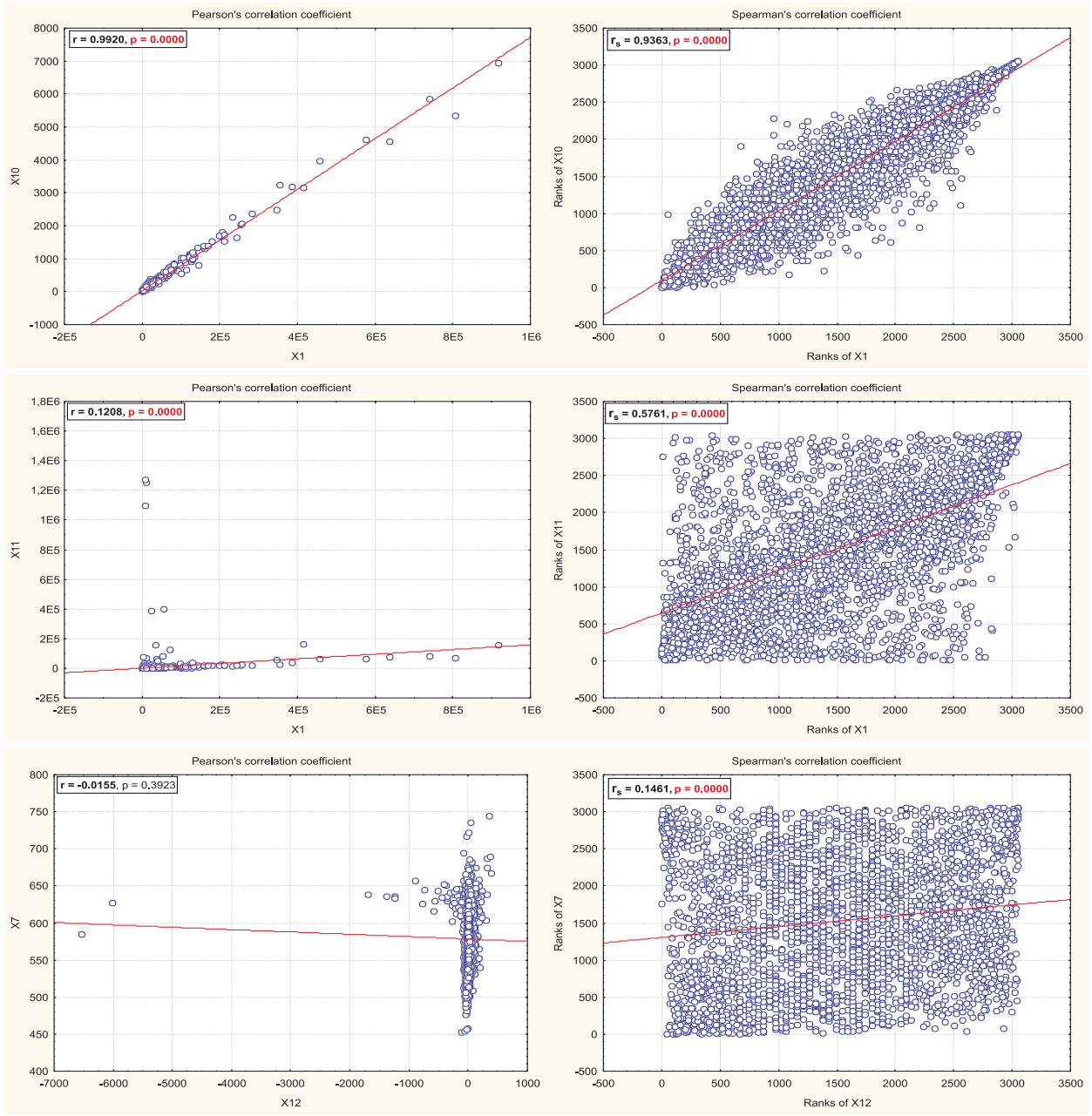

Fig. 3. Comparison of Pearson's and Spearman's coefficients of correlation of selected variables for communes (divided into their urban and rural parts).

we have an insignificant and negative Pearson's correlation, and a significant and positive Spearman's correlation.

3. Conclusion

When analysing both Pearson's and Spearman's coefficients, one could logically expect that the significance of one would imply the significance of the other. On the other hand, a reverse implication does not necessarily seem to be logically true. As we have shown in the previous paragraph, the significance of Spearman's correlation can lead to the significance or non-significance of Pearson's correlation coefficient even for big sets of data, which is consistent with a logical understanding of the difference between the two coefficients. However, the logical reasoning is not correct in the case of the significance of Pearson's coefficient translating into the significance of Spearman's coefficient. It is possible to meet

a situation where Pearson's coefficient is negative while Spearman's coefficient is positive.

The above leads to the following conclusion: *Make sure not to overinterpret Spearman's rank correlation coefficient as a significant measure of the strength of the associations between two variables*.

References

Anderson T.W., 1996. R.A. Fisher and multivariate analysis. *Statistical Science* 11 (1): 20–34.

Bravais A.,1846. Analyse mathématique sur les probabilités des erreurs de situation d'un point. *Mémoires présentés par divers savants à l'Académie Royale des Sciences de l'Institut de France* 9: 255–332.

Daniels H.E., 1944. The relation between measures of correlation in the universe of sample permutations. *Biometrika* 33 (2): 129–135.

Fisher R.A., 1915. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika* 10: 507–521.

Fisher R.A., 1921. On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron* 1: 3–32.

Galton F., 1869. Hereditary genius. An inquiry into its laws and consequences. MacMillan, London.

Galton F., 1875. Statistics by intercomparison. *Philosophical Magazine* 49: 33–46.

Galton F., 1885. Regression towards mediocrity in hereditary stature. *Journal of the Anthropological Institute* 15: 246–263.

Galton F., 1877. Typical laws of heredity. *Proceedings of the Royal Institution* 8: 282–301.

Galton F., 1888. Co-relations and their measurement, chiefly from anthropometric data. *Proceedings of the Royal Society of London* 45: 135–145.

Galton F., 1890. Kinship and correlation. *North American Review* 150: 419–431.

Griffith D.A., 2003. *Spatial autocorrelation and spatial filtering*. Springer, Berlin.

Haining R., 1991. Bivariate correlation with spatial data. *Geographical Analysis* 23 (3): 210–227.

Kendall M.G., 1938. A new measure of rank correlation. *Biometrika* 30: 81–89.

Kendall M.G.,1948. *Rank correlation methods*. 4th ed. Griffin, London.

Pearson K., 1896. Mathematical contributions to the theory of evolution. III. Regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society Ser. A* 187: 253–318.

Pearson K., 1900. Mathematical contributions to the theory of evolution. VII. On the correlation of characters not quantitatively measurable. *Philosophical Transactions of the Royal Society Ser. A* 195: 1–47.

Pearson K., 1920. Notes on the history of correlation. *Biometrika* 13: 25–45.

Piovani J.I., 2008. The historical construction of correlation as a conceptual and operative instrument for empirical research. *Quality & Quantity* 42: 757–777.

Plata S., 2006. A note on Fisher's correlation coefficient. *Applied Mathematical Letters* 19: 499–502.

Rodgers J.L. & Nicewander W.A., 1988. Thirteen ways to look at the correlation coefficient. *The American Statistician* 42 (1): 59–66.

Spearman C.E, 1904a. The proof and measurement of association between two things. *American Journal of Psychology* 15: 72–101.

Spearman C.E, 1904b. General intelligence, objectively determined and measured. *American Journal of Psychology* 15: 201–293.

Spearman C.E., 1910. Correlation calculated from faulty data. *British Journal of Psychology* 3: 271–295.

Stigler S.M., 1988. Francis Galton's account of the invention of correlation. *Statistical Science* 4 (2): 73–86.

Student, 1908. Probable error of a correlation coefficient. *Biometrika* 6: 302–310.

Valz P.D. & Thompson M.E., 1994. Exact inference for Kendall's S and Spearman's rho. *Journal of Computational and Graphical Statistics* 3: 459–472.

Walker H. M., 1928. The relation of Plana and Bravais to theory of correlation. *Isis* 10 (2): 466–484.

Weida F.M., 1927. On various conceptions of correlation. *The Annals of Mathematics, Second Series* 29 (1/4): 276–312.

Xu Weichao, Yunhe Hou, Hung Y.S. & Yuexian Zou, 2010. Comparison of Spearman's rho and Kendall's tau in normal and contaminated normal models. Manuscript submitted to IEEE Transactions on Information Theory (http://arxiv.org/PS_cache/arxiv/pdf/1011/1011.2009v1.pdf)

Yule G.U., 1897a. On the significance of Bravais' formulae for regression, in the case of skew correlation. *Proceedings of the Royal Society of London Ser. A* 60: 477–489.

Yule G.U., 1897b. On the theory of correlation. *Journal of the Royal Statistical Society Ser. A* 60: 812–854.

Yule G.U., 1903. Notes on the theory of association of attributes in statistics. *Biometrika* 2: 121–134.

Yule G.U., 1907. On the theory of correlation for any number of variables, treated by a new system of notation. *Proceedings of the Royal Society of London* 79: 182–193.