Modern Applied Science; Vol. 10, No. 7; 2016
ISSN 1913-1844 E-ISSN 1913-1852
Published by Canadian Center of Science and Education

Non-Traditional Correlation Analysis: Explanatory Power and Opportunities for Knowledge Discovery in Democracy Studies

Nikolay P. Tretyakov[1,2], Pavel E. Golosov[1] & Saif A. Mouhammad[3]

[1] Russian Presidential Academy of National Economy and Public Administration, Moscow, Russia

[2] Russian State Social University, Moscow, Russia

[3] Taif University, Taif, Saudi Arabia

Correspondence: Nikolay P. Tretyakov, Russian Presidential Academy of National Economy and Public Administration, Prospekt Vernadskogo, 82, Moscow, Russia. E-mail: trn11@rambler.ru

Received: December 10, 2015 Accepted: January 16, 2016 Online Published: April 28, 2016

doi:10.5539/mas.v10n7p99 URL: http://dx.doi.org/10.5539/mas.v10n7p99

Abstract

The possibility of successful applications of the modified correlation coefficient is demonstrated. The latter was proposed by Lukashin nearly twenty five years ago and has been unused since then. A multivariate generalization of this coefficient is proposed. The modified correlation coefficients provide an efficient tool to develop a new multivariate classification method, i.e. a technique for grouping of objects that occurs together with their ranking. As an example of application of the new method, the data of Freedom House is used. NCA (Non-traditional Correlation Analysis), along with similar unconventional methods as FCA (Formal Concept Analysis) and QCA (Qualitative Comparative Analysis) allow to gain additional knowledge from existing databases and numerous ratings which are produced by different agencies. The latters often lack time and opportunities to deeply analyze them, even to go beyond a simple "averaging". NCA may give additional opportunities for social researchers to understand social phenomena in its complexity, for in-depth analysis and interpretation of structure of data, to build "hierarchical typologies", and broadly, for data mining and additional knowledge discovery.

Keywords: democracy, correlation analysis, time series, multivariate classification, multivariate ranking, multivariate clustering

1. Introduction

Traditional analysis is often used for studying statistical relations between variables that are represented by time series (Anderson, 1971). Correlation analysis is on the one hand easy to use and on the other hand it allows the determination of the impact of various economic indicators. This is why it is quite popular among sociologists and economists.

Classical correlation theory has been developed for stochastic stationary processes that allow any number of observations. This theory was primarily intended for technical applications, where the hypothesis of stationarity is acceptable and there are no principal limits for the number of realizations. However, as noted by (Lukashin, 2003), this approach is viewed not critically by economists and is widely used to analyze processes that are knowingly non-stationary (for example with trend) and even non-stochastic, which are typically presented by a single realization. Obviously, numerous realizations under similar conditions are usually impossible in sociology and economics.

Postulation of steady-state conduction is a necessary measure to which economists resort to provide some opportunity to conduct correlation analysis for a pair of variables. Actually, variables x_1 and x_2 usually do not have fixed average levels, as well as any specific standard deviations from them. These values m_1 and m_2 represent some conditional levels about which deviations of the series are calculated. In this context the usual indicators of correlation more likely express relations between deviations from the averages than between the series themselves. What is more, if one moves the boundaries of the sampling period, then all of the evaluations of averages and standard deviations will change. This is especially true in regard to time series with trends. A solution for this difficult situation is usually found by converting the original non stationary series to an about stationary one. To this end, one usually attempts to exclude time-trends from the series. However, firstly, this operation also depends on the sampling period's boundaries and the choosing of the trend's type. Secondly, very significant information may be removed from the time series as a result of the elimination of the trend. Residuals, as a rule, are weakly correlated,

while the original series can exhibit high correlation due to similar trends. Hence the distortion of primary information is only getting worse with exception of trends, as demonstrated by Lukashin (2003).

2. Method

However, another approach to correlation is possible. It consists in evaluation of coincidence or not coincidence of signs of variables' increments. To get an overview of the average correlation attributes of two non-stationary series, the following modified correlation coefficient was proposed (Lukashin, 2003, 1992):

$$r_{\text{mod}} = \sum_{t=2}^{n} \Delta y_{1t} \Delta y_{2t} \Big/ \sum_{t=2}^{n} \left| \Delta y_{1t} \Delta_{2t} \right| \tag{1}$$

Here $\Delta y_{jt} = y_{jt} - y_{j(t-1)}$ and the denominator in equation (1) plays the role of a normalizing coefficient.

Thanks to it $-1 \le r_{\text{mod}} \le 1$. With this coefficient, the above-mentioned disadvantages are eliminated and hence no deformation of primary data occurs.

Although this coefficient was proposed nearly twenty five years ago, it has been unused since then. The aim of the present work is to demonstrate the efficiency of non-traditional correlation coefficients in case of the cumulative way of their calculation (i.e. they are calculated at each instant of time over cumulative sums of variables' increments).

A multivariate generalization of (1) may be as follows:

$$r_{\text{mod}} = \sum_{t=2}^{n} \Delta y_{1t} \Delta y_{2t} \ldots \Delta y_{mt} \Big/ \sum_{t=2}^{n} \left| \Delta y_{1t} \Delta_{2t} \ldots \Delta y_{mt} \right| \tag{2}$$

Another pssibility is to apply non-traditional correlation coefficients not to time series but to multidimensional data analysis. It turns out that the modified correlation coefficient provides an efficient tool to develop a new multivariate classification method. Indeed, a general question facing researchers in many areas of inquiry is how to organize observed data into meaningful structures, that is, to develop taxonomies (StatSoft, Inc., 2015). There are two main classification problems: grouping and ranking. Cluster analysis is an exploratory data analysis tool which aims at sorting different objects into groups, while dimensionality reduction techniques (e.g. principal component analysis) help to solve ordering problems. It would be very interesting to develop an unifying method for dealing with both of the problems. Such a procedure would allow, for instance, to produce grouping of objects that occurs in a natural way together with their ranking. We propose the following algorithm. Step 1. Choose a set of variables. Step 2. Rank the objects by the values of one of the variables. Step 3. Calculate the modified correlation coefficients using (1) for pairwise coefficients or (2) for multiple correlation (where *t* is not time but the objects' numbers in accordance to the ranking). Step 4. Analyse the plots of the coefficients and interpret the results.

An example of application is given below in Sec. 3. This calculation and numerous other examples show that the graphs of modified correlation coefficients (1) and (2) often demonstrate expressed stepwise structures. This can be interpreted as clustering of ranked objects. Interestingly, traditional correlation coefficients (Pearson and Spearman) very rarely demonstrate such a stepwise appearance and so they are less suitable for the analysis.

3. Example of Multivariate Data Analysis

As an example of use of the new method, let us refer to data of Freedom House, which is an independent watchdog organization dedicated to the expansion of freedom around the world (Freedom House, 2014). Nations in Transit is the comparative and multivariate study of reforms in the former Communist states of Europe and Eurasia. Nations in Transit tracks the reform record of 29 countries and provides Freedom House's data about this vast and important region. Countries are rated on a scale of 1 to 7, with 1 representing the highest and 7 the lowest level of democratic progress Note that of the 29 countries assessed in 2014, 14 (from Slovenia to Macedonia) were rated as stable democracies, 5 (from Albania to Ukraine) as transitional or hybrid political systems, and 10 (from Kosovo to Uzbekistan) as authoritarian regimes. In 2013, there was a more detailed classification: stable democracies (Slovenia, Estonia, Latvia, Poland, Czech Republic, Lithuania, Slovakia, Hungary), partial democracies (Bulgaria, Romania, Serbia, Croatia, Montenegro, Macedonia), transitional or hybrid political systems (Albania, Bosnia and Herzegovina, Georgia, Moldova, Ukraine), partially authoritarian political systems (Kosovo, Armenia, Kyrgyzstan), and authoritarian political systems (Russia, Tajikistan, Kazakhstan, Azerbaijan, Belarus, Turkmenistan, Uzbekistan) (Freedom House, 2014). This brings up the question: may such a grouping be obtained via any method?

Considering that DS is the average of all the ratings, we do not incorporate it into the set of variables for which the multivariate coefficient (2) is calculated. Figures 3 - 6 show the graphs of this coefficient for some different rankings of the countries. Note that the change of ranking leads to cardinal changes of the multiple coefficient's pattern. It follows from the formula (2), which contains the increments Δy_{jt} that depend on the objects' ranking (recall that t is not time here but the countries' numbers in accordance to the ranking).

One problem with the ranking is that there are many equal values in Table 1. This leads to the necessity of using of a second ranking variable. As such we take DS.

Table 1. Nations in transit 2014: ratings and democracy score summary (Freedom House, 2014). Categories: 1. Electoral Process (EP). 2. Civil Society (CS). 3. Independent Media (IM). 4. National Democratic Governance (NGOV). 5. Local Democratic Governance (LGOV). 6. Judicial Framework and Independence (JFI). 7. Corruption (CO). 8. The Democracy Score (DS)

	EP	*CS*	*IM*	*NGOV*	*LGOV*	*JFI*	*CO*	*DS*
Slovenia	1,5	2	2,25	2	1,5	1,75	2,5	1,93
Estonia	1,75	1,75	1,5	2,25	2,5	1,5	2,5	1,96
Latvia	1,75	1,75	2	2	2,25	1,75	3	2,07
Poland	1,25	1,5	2,5	2,5	1,5	2,5	3,5	2,18
Czech Republic	1,25	1,75	2,75	3	1,75	1,75	3,5	2,25
Lithuania	2	1,75	2,25	2,75	2,5	1,75	3,5	2,36
Slovakia	1,5	1,75	2,75	3	2,5	3	3,75	2,61
Hungary	2,25	2,25	3,5	3,75	2,75	2,5	3,75	2,96
Bulgaria	2,25	2,25	4	3,75	3	3,25	4,25	3,25
Romania	3	2,5	4,25	3,75	3	3,75	4	3,46
Serbia	3,25	2,25	4	3,75	3,5	4,5	4,25	3,64
Croatia	3,25	2,75	4	3,5	3,75	4,5	4	3,68
Montenegro	3,5	2,75	4,25	4,25	3,25	4	5	3,86
Macedonia	3,25	3,25	5	4,25	3,75	4,25	4,25	4
Albania	4	3	4	4,75	3,5	4,75	5,25	4,18
Bosnia and Herzegovina	3,25	3,5	4,75	5,75	4,75	4,25	4,75	4,43
Georgia	4,5	3,75	4	5,5	5,5	5	4,5	4,68
Moldova	4	3,25	5	5,5	5,75	4,75	5,75	4,86
Ukraine	4	2,5	4,25	6	5,5	6	6,25	4,93
Kosovo	4,75	3,75	5,75	5,5	4,75	5,5	6	5,14
Armenia	5,75	3,75	5,75	5,75	5,75	5,5	5,25	5,36
Kyrgyzstan	5,5	4,5	6	6,5	6,25	6,25	6,25	5,89
Russia	6,75	5,75	6,25	6,5	6	6	6,75	6,29
Tajikistan	6,75	6,25	6,25	6,5	6	6,25	6,25	6,32
Kazakhstan	6,75	6,5	6,75	6,75	6,5	6,5	6,5	6,61
Azerbaijan	7	6,5	6,75	6,75	6,5	6,5	6,75	6,68
Belarus	7	6,5	6,75	6,75	6,75	7	6,25	6,71
Turkmenistan	7	7	7	7	6,75	7	6,75	6,93
Uzbekistan	7	7	7	7	6,75	7	6,75	6,93

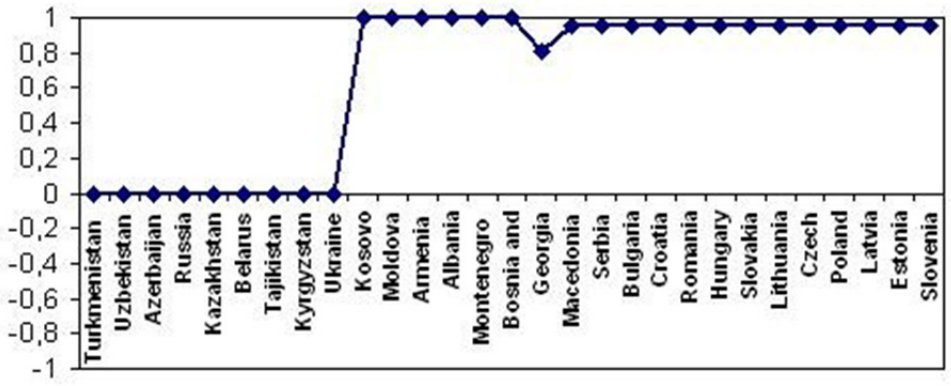

Figure 1. The multiple modified correlation coefficient, sorting by CO in descending order

As mentioned above, the graphs of coefficients (2) demonstrate horizontal flat intervals, and one may interpret respective points (objects) as belonging to the same cluster. Figure 1 shows the graph of the multiple coefficient calculated for the countries sorted by the parameter CO in descending order. As may be seen, it consists of three flat regions and this grouping practically coincides with the Freedom House's 2014 classification (with the exception of Montenegro which fell into the transitional group and Georgia that shifted close to the democratic group and occupied a special position there).

A very expressive graph is shown in Figure 2. It is the case of sorting by DS in descending order. The grouping practically coincides with the Freedom House's 2013 classification, only without separation of stable and partial democracies (though Slovenia, Latvia and Estonia are slightly separated from other countries, i.e. they may be regarded as the most democratic of all others). The transition from democracy to authoritarity justifies its name, being strongly oscillatory.

Figure 3 (sorting by NGOV in descending order) practically reproduces the 2013 classification too, without separation of hybrid and partially authoritarian categories (with the exception of Ukraine that fell into the authoritarian group and Hungary that moved slightly to the left and appeared in the beginning of the partially democratic region). Note that the special situation of Slovenia, Latvia and Estonia is more expressed here than in the former example. As noted above, our method produces grouping of objects that occurs together with their ranking. The parameter by which the ranking is made, sets a main subject of the classification. Clearly, different rankings will produce different groupings. In the former examples, we have successfully reproduced the Freedom House's groupings. In Figure 4, the graph corresponding to sorting by CS (Civil Society) in descending order is reproduced. The grouping is somehow different from the former cases (however, not crucially).

When analysing multivariate data, pairwise coefficients (1) may be useful too. However, they more rarely present such expressed flat horizontal regions as multidimensional coefficients do. By way of example, the correlation coefficient between JFI and DS is shown in Figure 5 (ranking by DS in ascending order). One may see that the usual Pearson's correlation coefficient does not provide any clustering. One more ranking (by NGOV in descending order) leads to the configuration represented in Figure 6. It looks as though there are two worlds with totally different regularities with a small buffer region (Armenia, Ukraine, Kyrgyzstan) between them.

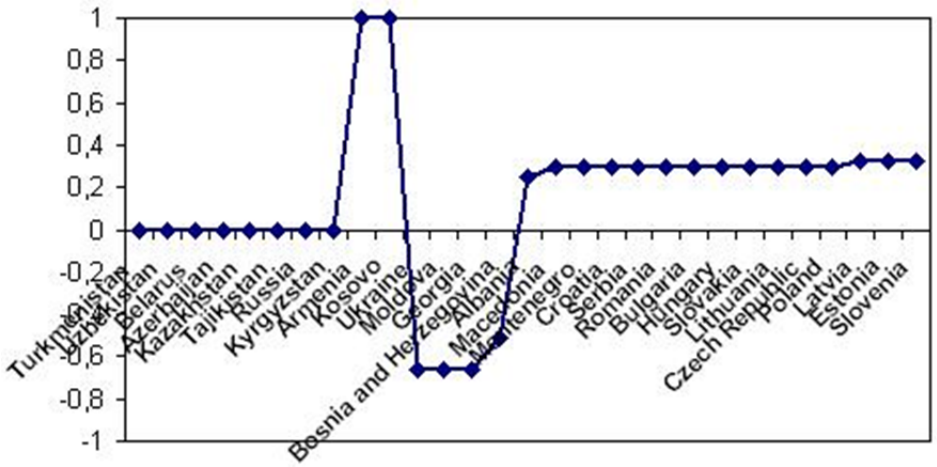

Figure 2. The multiple modified correlation coefficient, sorting by DC in descending order

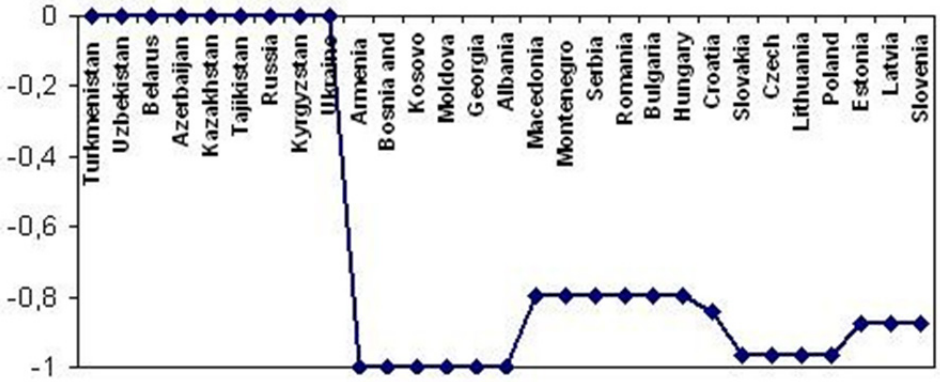

Figure 3. The multiple modified correlation coefficient, sorting by NGOV in descending order

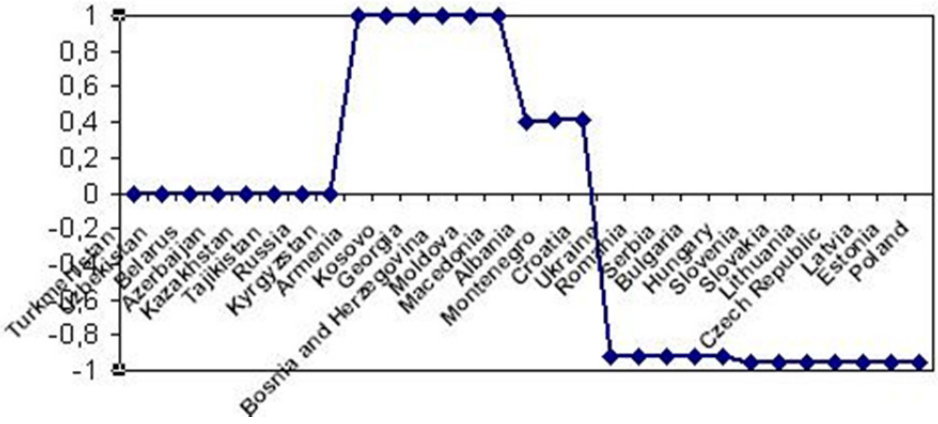

Figure 4. The multiple modified correlation coefficient, sorting by CS in descending order

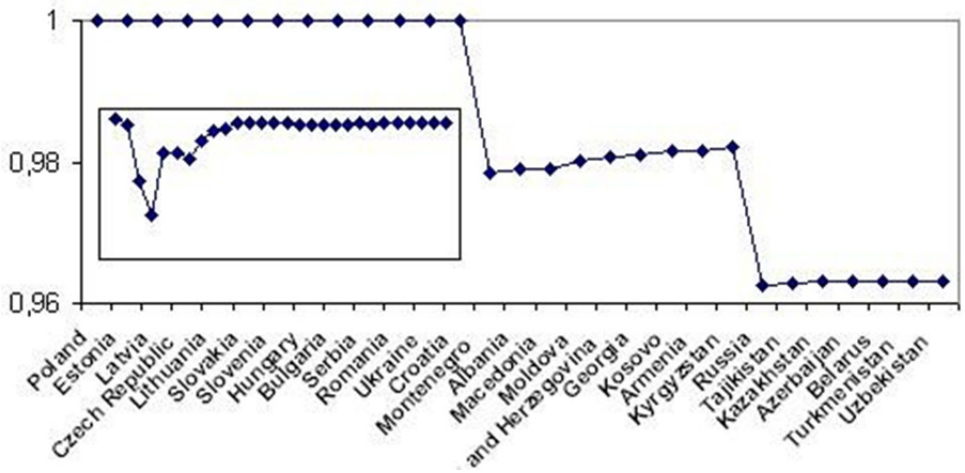

Figure 5. The pairwise modified correlation coefficient between JFI and DS, sorting by CS in ascending order. The Pearson's correlation coefficient is shown in the small window.

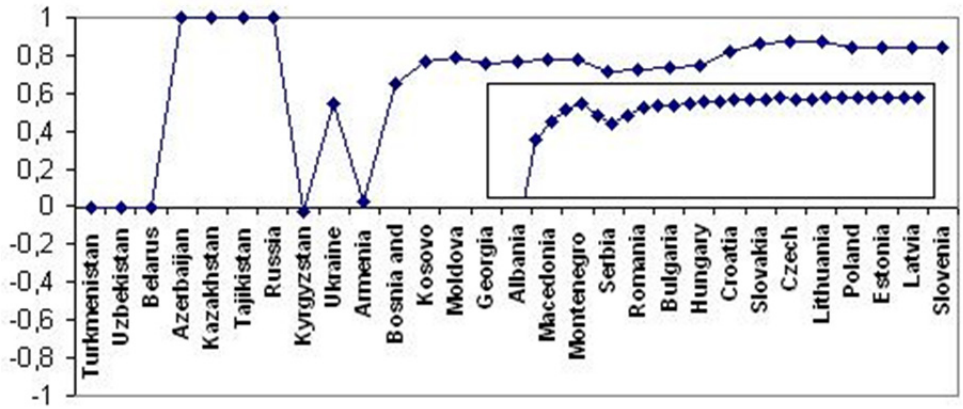

Figure 6. The pairwise modified correlation coefficient between JFI and DS, sorting by NGOV in descending order. The Pearson's correlation coefficient is shown in the small window

4. Conclusion

Correlations of time series are not constant but vary in time. Usual correlation coefficients are less suitable for their analysis than the modified correlation coefficient (1). We propose its multivariate generalization (2). Note that in statistics, the so-called "coefficient of multiple correlation" measures a strength of the relationship between one single variable and a set of other variables. On the contrary, all variables in (2) enter symmetrically. In other words, it may be considered as a real multiple correlation coefficient that measures the mean mutual relationship in a set of variables.

These modified correlation coefficients are very promising for multivariate classification too. We propose a new technique for grouping of objects that occurs together with their ranking. NCA, along with similar methods as FCA and QCA, allow to gain additional knowledge from existing databases and numerous ratings which in large quantities are produced by different agencies. The latters often lack time and opportunities to deeply analyze them, even to go beyond a simple "averaging". NCA may give additional opportunities for social researchers to understand social phenomena in their complexity, for in-depth analysis and interpretation of structure of data, to build "hierarchical typologies", and broadly, for data mining and additional knowledge discovery.

However, it is necessary to recognize that at present this practice partially corresponds to the famous phrase "It is more of an art than a science". Indeed, in order to obtain such successful results as, for instance, Figures 4, 5 or 7, one has to select an appropriate set of variables along with a suitable objects' ranking. It would be useful to develop algorithms capable to formalize such processes.

References

Anderson, T. W. (1971). *The Statistical Analysis of Time Series*. John Wiley Sons, Inc. Retrieved from http://onlinelibrary.wiley.com/book/10.1002/9781118186428

Campbell, J. Y., Lo, A. W., & MacKinlay, A. C. (1996). *The Econometrics of Financial Markets.* NJ: Princeton University Press. Retrieved from http://press.princeton.edu/chapters/s5904.pdf

Freedom House. (2014). Retrieved from https://freedomhouse.org/report/nations-transit/nations-transit-2014 https://freedomhouse.org/sites/default/ les/Data tables.pdf

Ganter, B., & Wille, R. (1999). *Formal Concept Analysis: Mathematical foundations*. Springer, Berlin. Retrieved from http://www.springer.com/us/book/9783540627715

Lukashin, Yu. P. (1992). Non-traditional correlation analysis of time series. *Economics and Mathematical Methods, 28*, 406-413 (in Russian).

Lukashin, Yu. P. (2003). Adaptive methods of short-term forecasting of time series. *Finance and statistics,* Moscow (in Russian).

Morlino, L. (2011). Changes for democracy: Actors, structures, processes. Oxford University Press, NY. Retrieved from https://global.oup.com/academic/product/changes-for-democracy-9780199698110?cc=us& lang=en&

Copyrights