STATISTICS IN TRANSITION new series, Spring 2014
Vol. 15, No. 2, pp. 309–316

TESTS FOR CONNECTION BETWEEN CLUSTERING OF POLISH COUNTIES AND PROVINCE STRUCTURE

Małgorzata Markowska[1], **Marek Sobolewski**[2], **Andrzej Sokołowski**[3], **Danuta Strahl**[4]

ABSTRACT

The general idea of statistical tests which allow testing the influence of geographical or administrative units of upper level on clustering results of lower level units is presented, basing on the authors' earlier works. The so-called "active border" notion is used in these methods. If two counties *(powiats)* have been classified into different clusters then the border between them is called active. This border can be also the border between provinces. The number and length of active borders are used in the proposed test statistics. Their distribution depends on the actual geographic division of a given country. In this paper we present results for Poland and division for provinces and counties. Tables for test critical values and the approximation functions are given.

Key words: cluster analysis, NUTS, comparing partitions.

1. Introduction

In Sokolowski *et al.* (2013a), (2013b) the general idea of statistical tests which allow testing the influence of geographical or administrative units of upper level on clustering results of lower level units was presented. The so-called "active border" notion is used in these methods. If two counties have been classified into different clusters then the border between them is called active. This border can be also the border between provinces. If the upper level has no influence on the lower level partition results, then only randomness should decide if the active border is also the upper level border. Distribution of test statistic depends on the actual geographic division of a given country. In this paper we present results for Poland and division for provinces and counties. There are 978 borders between counties and 210 of them are also elements of between-province

[1] Wroclaw University of Economics. E-mail: malgorzata.markowska@ue.wroc.pl.
[2] Rzeszow University of Technology. E-mail: mareksobol@poczta.onet.pl.
[3] Cracow University of Economics. E-mail: andrzej.sokolowski@uek.krakow.pl.
[4] Wroclaw University of Economics. E-mail: danuta.strahl@ue.wroc.pl.

border. The total length of borders between counties sums up to 88,869 km, including 16,062 km of borders between provinces.

2. Test statistics

Both of the proposed test statistics are being used for testing the same following hypotheses:

H_0: Province level has no influence on the partition results obtained for counties

H_1: Province level influences significantly the results of counties partition

It seems natural that if the upper level has no influence on the lower level partition results, then only randomness should decide if the active border is also the upper level border. If "too many" active borders between counties are also borders between provinces one should reject the null hypothesis. Thus, the proposed tests have right-sided critical region. The following two test statistics are considered:

$$L1 = \frac{Number\ of\ active\ borders\ which\ are\ borders\ between\ provinces}{Total\ number\ of\ observed\ active\ borders} \tag{1}$$

$$L2 = \frac{Length\ of\ active\ borders\ which\ are\ borders\ between\ provinces}{Total\ length\ of\ observed\ active\ borders} \tag{2}$$

3. Simulation study

For partition simulations we use the simplified version (with given number of groups) of random partition generator proposed by Sokolowski (1979):

- set *k* (number of groups),
- assign random number from uniform distribution to each object,
- order objects according to values of this random variable,
- now we have *(n-1)* potential borders between objects,
- assign random number from uniform distribution to each potential border,
- make "active" first *k* borders with the biggest values of these random numbers.

We have considered partitions of Poland's counties from 2 to 16 groups. With 1000 simulation runs we have found that the distribution of *L1* and *L2* statistics can be approximated by normal distribution. Empirical distribution for *k=5* as an example is presented on Fig 1.

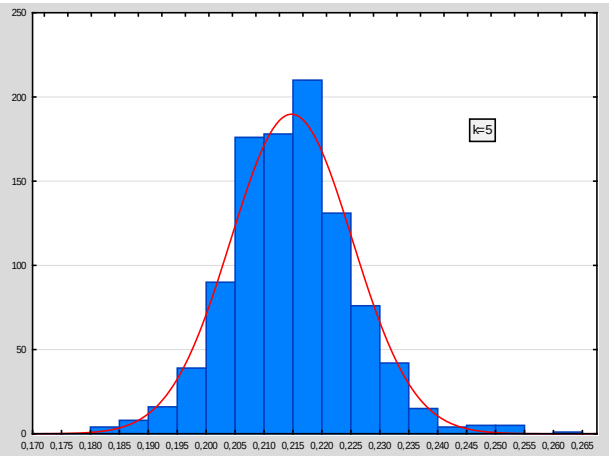

Figure 1. Empirical distribution of L1 statistic under null

Expected value of L1 equals 210/978=0.2147, while standard distribution depends on the number of clusters, but it can be very well approximated by (1), see Fig.2

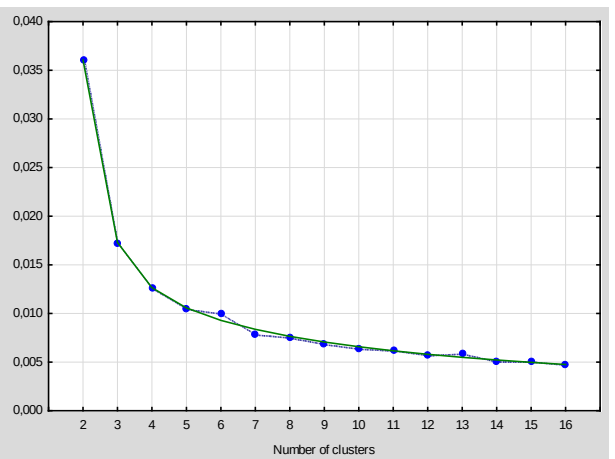

Figure 2. Standard deviation of L1 depending on the number of clusters

Critical values for 0.05 and 0.10 significance levels can also be approximated by fractional polynomials. Fig. 3 gives just one example of goodness-of-fit.

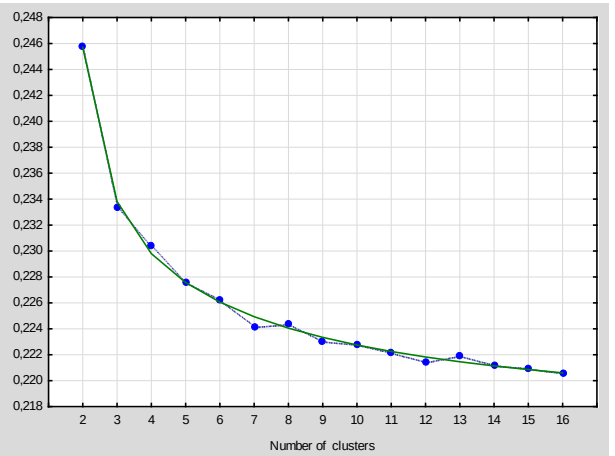

Figure 3. Critical value of L1 approximation for α=0.10

We have found that the distribution of L2 statistic can be also approximated by normal distribution. Expected value of L2 under null equals 16062/88869=0.1807 while standard deviation and critical values follows well fitted functions. Approximating functions are given in Table 1 and smoothed critical values in Table 2.

Table 1. Approximating functions

Parameter	Function	Adjusted coefficient of determination	Standard error of residuals
SD(L1)	$0.000802+0.074792k^{-1}-0.210665k^{-2}+0.404299k^{-3}$	0.999	0.0003
$Q_{0.90}$(L1)	$0.216349+0.076325k^{-1}-0.145002k^{-2}+0.219535k^{-3}$	0.996	0.0004
$Q_{0.95}$(L1)	$0.215771+0.119359k^{-1}-0.289479k^{-2}+0.446677k^{-3}$	0.995	0.0006
SD(L2)	$0.002043+0.052128k^{-1}-0.098926k^{-2}+0.230704k^{-3}$	0.999	0.0002
$Q_{0.90}$(L2)	$0.184717+0.037557k^{-1}+0.032073k^{-2}$	0.996	0.0004
$Q_{0.95}$(L2)	$0.184124+0.080401k^{-1}-0.121491k^{-2}+0.233468k^{-3}$	0.997	0.0005

Table 2. Critical values

Number of clusters	L1		L2	
	α=0.10	α=0.05	α=0.10	α=0.05
2	0.2457	0.2589	0.2115	0.2231
3	0.2338	0.2399	0.2008	0.2061
4	0.2298	0.2345	0.1961	0.2003
5	0.2276	0.2316	0.1935	0.1972
6	0.2261	0.2297	0.1919	0.1952
7	0.2249	0.2282	0.1907	0.1938
8	0.2241	0.2270	0.1899	0.1927
9	0.2233	0.2261	0.1893	0.1919
10	0.2228	0.2253	0.1888	0.1912
11	0.2223	0.2246	0.1884	0.1906
12	0.2218	0.2240	0.1881	0.1901
13	0.2215	0.2234	0.1878	0.1897
14	0.2211	0.2230	0.1876	0.1893
15	0.2209	0.2226	0.1874	0.1890
16	0.2206	0.2222	0.1872	0.1887

4. Example

We have taken four variables characterizing Polish counties: number of births per 1000 population, unemployment rate, average salary and number of new flats per 1000 population. Ward's agglomerative clustering method suggests the division into seven clusters (see Fig. 4)

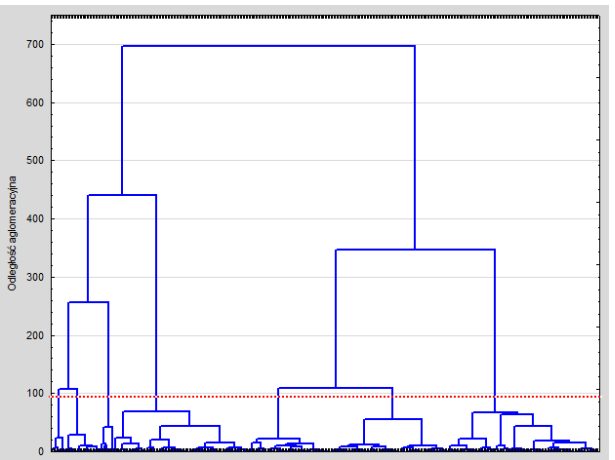

Figure 4. Ward's dendrogram of Polish counties

On Fig. 5 we can see the geographical distribution of clusters together with borders between provinces.

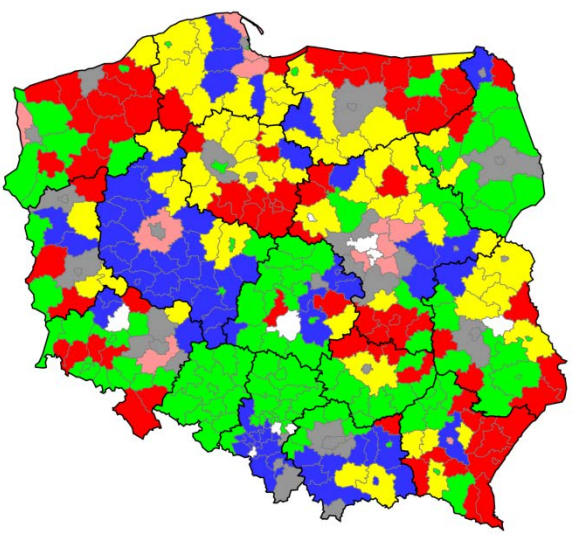

Figure 5. Clusters of Polish counties

Observed value of L1 equals 0.215 while critical value for α=0.10 is 0.2249 (p=0.450), and L2=0.185 with critical value 0.1907 and p=0.273. It is clear from both test statistics that there is no proof for statistically significant influence of province level on counties partition based on four considered variables.

5. Conclusions

It has been found that critical values for both proposed test statistics can be very well approximated by relatively simple functions while testing the influence of voivodship level of Polish provinces on county level. The example provided is only an illustrative effort. The proposed test can be widely used in testing the relations between administrative levels in Poland with respect to economic phenomena, politics, public administration and quality of life.

Acknowledgements

The project has been financed by the Polish National Centre for Science, decision DEC-2013/09/B/HS4/00509.

REFERENCES

SOKOŁOWSKI, A., (1979). Generowanie losowego podziału zbioru skończonego. Prace Naukowe Akademii Ekonomicznej we Wrocławiu [Generating random distribution of a finite set. Scientific Papers of Wrocław University of Economics], 1979, 160 (182), 413–415.

SOKOŁOWSKI, A., STRAHL, D., MARKOWSKA, M., SOBOLEWSKI, M., (2013a). The influence of upper level NUTS on lower level classification of EU regions. European Conference on Data Analysis, Luxembourg, July 10–12, 2013.

SOKOŁOWSKI, A., STRAHL, D., MARKOWSKA, M., SOBOLEWSKI, M., (2013b). The hierarchy test of geographic units based on border lengths, Conference of the International Federation of Classification Societies IFCS 2013, Tilburg, Netherlands, July 14–17, 2013. Abstract in Program and Book of Abstracts, Tilburg University, 46.