

NULL DISTRIBUTION OF MORAN’S *I* BY STRATIFICATION

Solymosi, N.^{1*}, Reiczigel, J.¹, Harnos, A.¹, Rubel, F.² and Berke, O.³

1 Introduction

Moran’s *I* for spatial autocorrelation is a frequently used test for spatial clustering. It is a weighted product-moment correlation coefficient, where the weights reflect to geographic proximity.

$$I = m \frac{\sum_{i=1}^m \sum_{j=1}^m w_{ij} (x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^m (x_i - \bar{x})^2 \sum \sum_{i \neq j} w_{ij}} \tag{1}$$

where x_i is the variable of interest, m is the number of districts (observational units) and

$$w_{ij} = \begin{cases} 1 & \text{if district } i \text{ and } j \text{ have a common boundary} \\ 0 & \text{if district } i \text{ and } j \text{ have not any common boundary.} \end{cases}$$

In some studies Moran’s *I* is calculated directly from observed counts [1, 2], while it is argued that it is valid only if calculated from rates [3, 4].

The conventionally used null distributions for Moran’s *I* are the normal distribution or distributions based on randomization or Monte Carlo [5]. All these assume that the distribution of the variable of interest is the same in all units.

In a former study we analyzed the pattern of the reported rabies cases in Hungary between 1990 and 2001, and in this case we did not have any reliable data of the fox population for calculating rates. As size of district and frequency showed strong correlation in each of the 3 parts of Hungary (Table 1, Figure 1), we tried to standardize the frequencies by the district size. But one can never know if such a standardization mirrors the true nature of the relationship. We think that an incorrectly chosen standardization - say when the true relationship is power or logarithmic, and standardization is linear - may not help much (or may make things even worse), therefore we aimed to find an alternative method, and test it on simulated data.

Table 1. Correlation between district size and frequency

Part of Hungary	R
Transdanube	0.629
Between the Danube and Tisza	0.585
Transtisza	0.761

Table 2. Descriptives of area structures

Measure	Transdanube	Between the Danube and Tisza	Transtisza
Area (km^2)	37137.89	31428.09	23754.37
Number of districts	1736	909	464
Number of nonzero links	10322	5224	2550

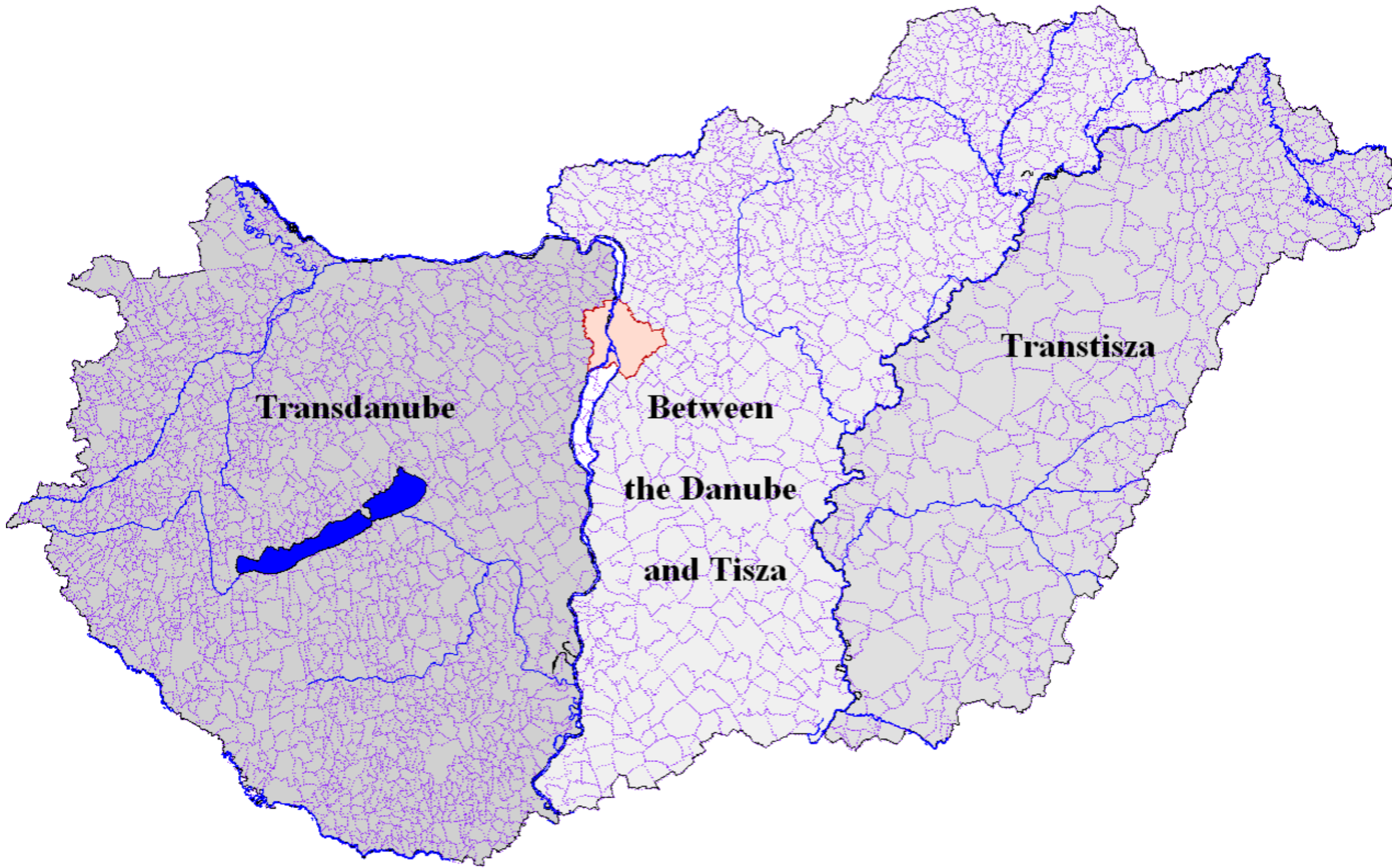


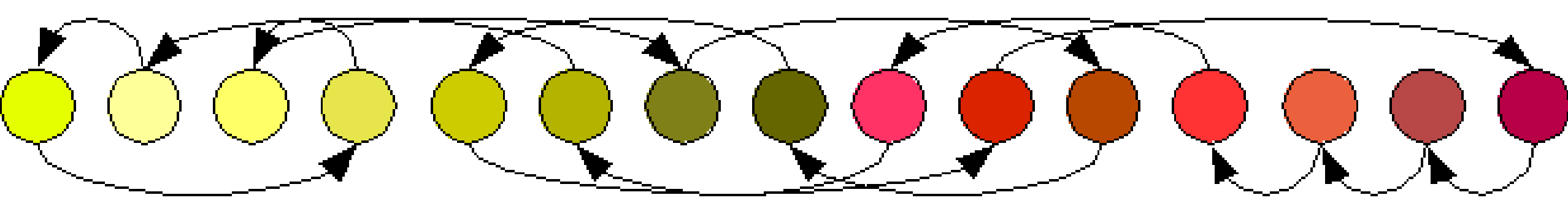
Figure 1. Map of Hungary

2 Material and method

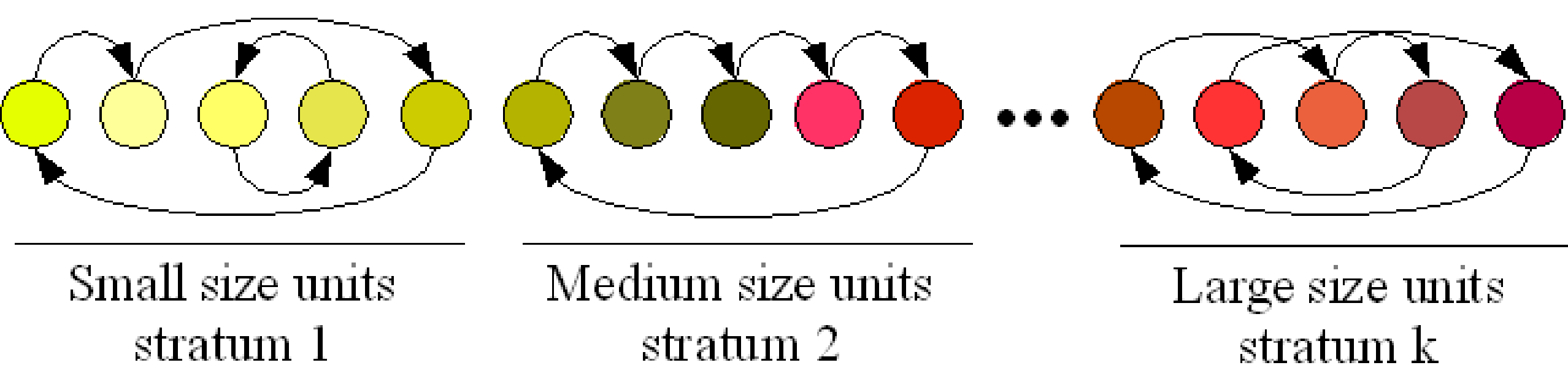
2.1 Monte Carlo method

In the Monte Carlo method, the null distribution is obtained by calculating Morans I for several simulated data sets, in which observed data (counts or whatever) are randomly assigned to the districts. Random assignment is made by randomly permuting the data values.

2.1.1 Conventional Monte Carlo method (permuting the whole dataset)



2.1.2 Stratified Monte Carlo method (permuting within strata)



2.2 Simulations

In the simulation we used spatially random *Poisson* distributed variables with λ proportional to the district size. Generated number of cases in district i is $Poisson(\lambda_i)$:

$$\lambda_i = \frac{\sum n_i}{\sum \sqrt{a_i}} \sqrt{a_i} \tag{2}$$

where n_i is the number of cases in the district i and a_i is the size of district i . We made 1000 simulations. In both Monte Carlo methods

¹Szent István University, Budapest, Hungary
²University of Veterinary Medicine, Vienna, Austria
³University of Guelph, Guelph, Ontario, Canada
School of Veterinary Medicine Hannover, Germany

we made 999 permutations. In the stratified method the length of each stratum was 30 units.

2.3 Software

As geographical data source a civil vector graphical digital map of Hungary (OTAB1) was used with a resolution of 1:100 000 (see Table 2 for further details). We used the *R* statistical software [6] (and in particular the packages *pps* and *spdep*) for all data analysis and simulation in cooperation with *VetEpiGIS* software (developed by the first author).

3 Results

Table 3. Type I error rates				
Method	x_i	Region		
		Transdanube	Between the Danube and Tisza	Transtisza
Randomization				
	n_i	0.069	0.192	0.363
	n_i/a_i	0.068	0.109	0.132
	$n_i/\sqrt{a_i}$	0.067	0.052	0.060
Monte Carlo				
	n_i	0.061	0.176	0.347
	n_i/a_i	0.057	0.093	0.106
	$n_i/\sqrt{a_i}$	0.061	0.045	0.050
Stratified MC				
	n_i	0.028	0.046	0.056

Here standardization n_i/a_i illustrates the effect of incorrectly chosen standardization on the results, while $n_i/\sqrt{a_i}$ represents the correct standardization (that corresponding to the true model).

4 Discussion & Conclusion

If the variable of interest has high correlation with the size of observational units or population then it is a good choice to stratify the recordset to homogeneous subsets as the base of the Monte Carlo permutation.

Here the subset size was arbitrary. Further research may focus on developing a method to define homogeneous strata. For example to reduce the correlation within the subsets by spalling.

Acknowledgements

This work was partly supported by the grant OTKA T035150, T032437 and ACTION Austrian-Hungarian Foundation 2003-2004 (53ÖU2).

References

[1] Weiskampel, J. F., Bagley, D. A., Ehrhart, L. M. and Rodenbeck, B. L., Spatiotemporal patterns of annual sea turtle nesting behaviors along an East Central Florida beach., *Biological Conservation*, , No. 110, pp. 295–303, 2003.
[2] Dale, M. R. T., Dixon, P., Fortin, M.-J., Legendre, P., Myers, D. E. and Rosenberg, M. S., Conceptual and mathematical relationships among methods for spatial analysis., *Ecography*, , No. 25, pp. 558–577, 2002.
[3] Carpenter, T. E., Methods to investigate spatial and temporal clustering in veterinary epidemiology., *Preventive Veterinary Medicine*, , No. 48, pp. 303–320, 2001.
[4] Pfeiffer, D. U., Spatial analysis - a new challenge for veterinary epidemiologist., *Proceedings of Annual Meeting of Society for Veterinary Epidemiology and Preventive Medicine*, edited by Thrushfield, M. V. and Goodall, E. A., pp. 86–106, Society for Veterinary Epidemiology and Preventive Medicine, March 2000.
[5] Cliff, A. D. and Ord, J. K., Spatial processes, pp. 63–65, Pion, 1981.
[6] R Development Core Team, R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, 2003, ISBN 3-900051-00-3.

*E-mail address: solymosinorbert@freemail.hu