ACTA CLIMATOLOGICA ET CHOROLOGICA
Universitatis Szegediensis, Tomus 40-41, 2007, 79-89.

OBJECTIVE ANALYSIS AND RANKING OF HUNGARIAN CITIES, WITH DIFFERENT CLASSIFICATION TECHNIQUES, PART 1: METHODOLOGY

L. MAKRA and Z. SÜMEGHY

Department of Climatology and Landscape Ecology, University of Szeged, P.O.Box 653, 6701 Szeged, Hungary
E-mail: makra@geo.u-szeged.hu

Összefoglalás – A tanulmány célja, hogy a magyarországi városokat és megyéket környezetminőségük és környezeti tudatosságuk szintje alapján osztályozza. Ahhoz, hogy ezt a feladatot megoldjuk, kiszámítottuk a „Green Cities Index", illetve a „Green Counties Index" értékeket, melyek alapján a városokat és a megyéket 7 különböző kategória 19 környezeti indikátora segítségével rangsoroltuk. Ezt követően azt a célt túztük ki, hogy összehasonlítsuk a különböző clusterező eljárásokat a városok és megyék osztályozásában. Az SPSS szoftver segítségével elvégzett clusteranalízis mind a városokra, mind a megyékre 6-6 homogén csoportot eredményezett. Az R-nyelv segítségével végrehajtott clusteranalízis az *agnes*, a *fanny* és a *pam* algoritmusok felhasználásával történt.

Summary – The aim of the study was to rank and classify Hungarian cities and counties according to their environmental quality and level of environmental awareness. To accomplish this task, „Green Cities Index" and „Green Counties Index" were calculated that rank cities and counties on the basis of seven different categories of 19 environmental indicators. Furthermore, our aim was to compare different methods in classifying cities and counties. Cluster analysis using SPSS software resulted in 6 homogenous groups for both the cities and the counties. Clustering with R-language was carried out using algorithms *agnes*, *fanny* and *pam*.

Key words: environmental indicators, Green Cities Index, Green Counties Index, factor analysis, clustering, SPSS-software, R-language; algorithms: agnes, fanny, pam

1. INTRODUCTION

In Hungary, 236 cities accounting for 65.7 % of the country's population were registered on January 1, 2001. Environmental factors in cities such as housing, transportation, air quality and public green space, etc., are important for the quality of life (*Kerényi*, 1995). But which cities have cleaner air, more urban parkland, or more pleasant climate? Which do a better job at organising traffic systems, waste management or public sanitation? Which cities are wasteful in their use of water or energy? To answer these questions, at least at a preliminary level, the so-called "Green Cities Index", which ranks cities on several environmental criteria, was developed (*Cutter*, 1992).

In this study, 25 environmental indicators were initially considered for each of the 236 Hungarian cities. Indicators, which were not available for all cities, were subsequently omitted. The cities were ranked by population and population density as well. However, these two parameters were not included for ranking according to the Green Cities Index, since larger and more densely populated cities do not necessarily have poorer environmental quality. Because environmental regulations in many cities have become increasingly more stringent, part of the data used in this study may be out of date by the date of publication. Consequently, Green Cities Index rankings should be viewed as a measure of environmental quality and concern at a given time.

The data basis for the study is drawn partly from the statistical yearbooks of Hungarian counties and Budapest for the year 2000 (*HCSO*, 2000a, 2000b) and partly from *Vaskövi* (2000).

2. OBJECTIVES

The aim of the study is to rank and classify Hungarian cities and counties according to their environmental quality and level of environmental awareness. A further aim of the study is to compare results received after performing hierarchical, agglomerative clustering techniques of both the SPSS software and the R-language. Besides, the spatial distribution of clusters received is also compared and their connection with the planning-statistical regions of Hungary is evaluated.

The different clustering techniques are found in most statistical softwares. These algorithms are represented by using the R-language.

3. METHODS

3.1. Environmental indicators

Seven different categories of environmental indicators ranging from water consumption to air quality were included in the Green Cities Index. Specific measures within each category were selected on the basis of data availability. Some related measures were combined to yield new, composite measures. Altogether 25 indicators were considered initially but only 19 were retained with 7 categories and their 19 indicator elements (*Table 1*). In *Table 1*, air quality is based on the average of the non-heating half-year (April 1, 2000 – September 30, 2000) and the average of the heating half-year (October 1, 2000 – March 31, 2001). Heating (cooling) degree-days are defined as the number of days when the mean temperature is above (below) 18°C, with each day weighted by the number of degrees above (below) 18°C. This parameter can be used as a measure of energy use for space heating (cooling) (Cutter, 1992). 18°C is considered the optimum temperature.

Data on all 19 indicators are available for only 88 of the 236 cities in the data base. Hence, further analyses are based on those 88. Though these indicators are neither perfect nor exhaustive, they enable an overall comparison among the relevant cities.

3.1.1. The Green Cities Index

The Green City Index is derived as follows

(a) The statistics for each indicator for each city were compiled from the yearbooks.

(b) Each indicator element was represented with a serial number (1 – 19).

(c) For each indicator element, cities were ranked from the most environmentally friendly (1) to least friendly (88) based on their statistics as determined in step (a). These ranks represent city scores on each indicator.

(d) The rank scores achieved by each city over the 19 indicator elements were averaged. The resulting figure is the Green City Index.

(e) Finally, the Green City Indices were ranked to yield the Final Sequence. The Final Sequence (FS) places the cities in rank order from the best (1) to the worst (88) based on step (d). FS is a rank of ranks.

Table 1 Categories and indicators used for compiling the Green Cities Index for the cities and counties

Categories	**Indicators**		
	Serial No.	**Elements**	**Units**
Water Consumption	1	Water use	m^3 / capita / year
Energy Consumption	2	Gas consumption	m^3 / household / year
	3	Electric energy consumption	kWh / household / year
	4	Degree days	sum of heating and cooling degree days
Public Utilities Supply	5	Ratio of households connected to gas network	percent
	6	Ratio of dwellings connected to drinking water network	percent
	7	Ratio of dwellings connected to public sewage system	percent
	8	Public sewage system	m / km drinking water conduit
Traffic	9	car supply	inhabitants per car
Waste management	10	Total drained-off waste water	m^3 / capita / year
	11	Total waste removed	m^3 / capita / year
	12	Ratio of dwellings connected to regular waste removal system	percent
Settlement amenities factors	13	Public green area	m^2 / capita
	14	Ratio of constructed inner roads	percent
	15	Ratio of constructed public surfaces cleaned regularly	percent
	16	Housing	occupants / dwelling

Table 1 (cont.)

Categories	Indicators		
	Serial No.	Elements	Units
Air Quality*	17	Average concentration of particulates deposited	g / m^2 / 30 days
	18	Average concentration of sulphur-dioxide	μg / m^3
	19	Average concentration of nitrogen-dioxide	μg / m^3

It is to be noted that the indicator elements were not weighted to reflect their relative importance to environmental quality or overall contribution to making a city liveable. Rather, they illustrate how each city fares when compared to others.

Human activities are the greatest source of contaminants in the environment. Thus, population and population density might be important environmental factors. But their implications to environmental quality are frequently contradictory since increases in the size of either variables or both do not automatically indicate a tendency towards poorer environmental quality. For example, compact and highly centralised cities with high population densities have the advantage of decreasing passenger car traffic between the city centre and the suburbs thus contributing to lower air pollution loads. However, such advantage may be countered by more concentrated sources of pollution and waste, and more congestion. On the other hand, cities that sprawl and are dispersed, resulting in lower population densities, may have a difficulty providing mass transit, but they may have more open space. On balance, large centralised cities tend to have greater difficulty achieving the same level of environmental quality than smaller cities. To test the impact of population and population density on the Green Index, a second set of Final Sequence (modified sequence) based on 21 indicator elements – the nineteen original ones, plus population and population density – was derived.

3.1.2. The Green Counties Index

The 19 Hungarian counties were also ranked from the most environment- friendly to the worst. The same environmental indicators as the ones used for the cities were applied. The so-called Green Counties Index values are the average of the scores achieved by all cities within the county. The Green Counties Index, similar to the Green Cities Index, is effectively a rank of ranks. Low numbers indicate better environmental quality.

3.2. Factor analysis

In order to reduce the dimensionality of the above-mentioned meteorological data sets and thus to explain the relations among the 19 variables (environmental indicators), the multivariate statistical method of factor analysis is used. The main object of factor analysis is to describe the initial variables $X_1, X_2, \ldots, X_p$ in terms of *m* linearly independent indices ($m < p$), the so called factors, measuring different *"dimensions"* of the initial data set. Each variable *X* can be expressed as a linear function of the *m* factors, which are the main contributors to the climate of Szeged:

$$X_i = \sum_{j=1}^{m} \alpha_{ij} F_j \qquad (1)$$

where α_{ij} are constant called factor loadings. The square of α_{ij} represents the part of the variance of X_i that is accounted for by the factor F_j.

One important stage of this method is the decision for the number (*m*) of the retained factors. On this matter, many criteria have been proposed. In some studies, the *Guttmann criterion* or *Rule 1* is used, which determines to keep the factors with eigenvalues > 1 and neglect those ones that do not account for at least the variance of one standardised variable X_i. Perhaps the most common method is to specify a least percentage (80 % in this paper) of the total variance in the original variables that has to be achieved (*Jolliffe*, 1993; *Sindosi et al.*, 2003). Extraction was performed by *Principal Component Analysis* (*k*th eigenvalue is the variance of the *k*th principal component). There is an infinite number of equations alternative to Eq. 1. In order to select the best or the desirable ones, the so-called "*factor rotation*" is applied, a process, which either maximises or minimises factor loadings for a better interpretation of the results. In this study, the "*varimax*" or "*orthogonal factor rotation"* is applied, which keeps the factors uncorrelated (*Jolliffe*, 1990, 1993; *Bartzokas and Metaxas*, 1993, 1995).

Factor analysis was applied on the tables of the initial data consisting, in the first case, of 19 columns (environmental indicators) and 88 rows for cities and, in the second, the same 19 columns (environmental indicators) and 19 rows for counties.

3.3. Cluster analysis

Clustering is an organizational methodology dating back to the Ancient Greeks. Aristotle was the first great classifier. He attempted to understand the essence of subgroups of the population. Observing that dolphins have a placenta, Aristotle reasoned that dolphins are mammals, not fish. This insight was greeted with almost uniform derision for nearly two thousand years. The fortunes of taxonomists have barely improved in the interim (*Gould*, 1996).

Objective classification of the cities and counties examined was achieved with the help of cluster analysis. The aim was to group cities and counties objectively based on their similarity in environmental conditions. The basis for the classification is to maximise the homogeneity of cities and counties within the clusters and maximise the heterogeneity among them. The database for the analysis consisted of city (county) scores in each of the 19 environmental indicators measured in 2000.

Cluster analysis is applied to the factor scores time series in order to objectively group days with similar weather conditions. The aim of the method is to maximize the homogeneity of objects within the clusters and also to maximize the heterogeneity between the clusters. Each observation (day) corresponds to a point in the *m*-dimensional space and each cluster consists of those observations, which are *"close"* to each other in this space.

3.3.1. Grouping procedures in R-language

Generally speaking, cluster analysis methods are of either of two types:

Partitioning methods: algorithms that divide the dataset into *k* clusters, where the integer *k* needs to be specified by the user. Typically, the user runs the algorithm for a range of *k*-values. For each *k*, the algorithm carries out the clustering and also yields a *„quality index"*, which allows the user to select a value of *k* afterwards.

Hierarchical methods: algorithms yielding an entire hierarchy of clustering of the dataset. Agglomerative methods start with the situation where each object in the dataset forms its own little cluster, and then successively merge clusters until only one large cluster remains which is the whole dataset. Divisive methods start by considering the whole dataset as one cluster, and then split up clusters until each object is separate.

Algorithms *Pam* and *fanny* of the partitioning type as well as algorithms *Agnes* and *mona* of the hierarchical type are considered in this paper.

3.3.1.1. Partitioning methods

Partitioning Around Medoids: function „pam"

The function *pam* is based on the search for *k* representative objects, called medoids, among the objects of the dataset (*Kaufman and Rousseeuw*, 1990). These medoids are computed so that the total dissimilarity of all objects to their nearest medoid is minimal: i.e. the goal is to find a subset $\{m_1,\ldots,m_k\} \subset \{1,\ldots,n\}$ which minimizes the objective function:

$$\sum_{i=1}^{n} \min_{t=1,\ldots,k} d(i,m_t) \cdot \tag{2}$$

Each object is then assigned to the cluster corresponding to the nearest medoid. That is, object *i* is put into cluster v_i when medoid m_{v_i} is nearer to *i* than any other medoid m_w, or

$$d(i,m_{v_i}) \le d(i,m_w) \text{ for all } w = 1,\ldots,k \; . \tag{3}$$

Finally *pam* provides a novel graphical display, the *silhouette plot* (*Rousseeuw*, 1986), and a corresponding *quality index* allowing to select the number of clusters. Let us first explain the silhouette plot. For each object *i* we denote by *A* the cluster to which it belongs, and compute

$$a(i) := \frac{1}{|A|-1} \sum_{j\in A, j\neq i} d(i,j) \, , \tag{4}$$

namely, *a(i)* measures average dissimilarity of *i* to all other objects of *A*.

Now consider any cluster *C* different from *A* and put

$$d(i,C) := \frac{1}{|C|} \sum_{j\in C} d(i,j) \, , \tag{5}$$

namely, *d(i,C)* measures average dissimilarity of *i* to all other objects of *C*.

After computing *d(i;C)* for all clusters $C \neq A$, we take the smallest of those:

$$b(i) := \min_{C \neq A} d(i, C). \tag{6}$$

The cluster *B* which attains this minimum [that is, *d(i;B)* = *b(i)*] is called the neighbour of object *i*. This is the second-best cluster for object *i*.

The silhouette value *s(i)* of the object *i* is defined as:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}. \tag{7}$$

Clearly, for *s(i)*: $-1 \le s(i) \le 1$. The value *s(i)* may be interpreted as follows:

s(i) ≈ *1* ⇒ object *i* is well classified (in *A*);
s(i) ≈ *0* ⇒ object *i* lies intermediate between two clusters (*A* and *B*);
s(i) ≈ *-1* ⇒ object *i* is badly classified (closer to *B* than to *A*).

The silhouette of the cluster *A* is a plot of all its *s(i)*, ranked in decreasing order. The entire silhouette plot shows the silhouettes of all clusters below each other, so the quality of the clusters can be compared: a wide (dark) silhouette is better than a narrow one.

The quality index mentioned earlier is the overall average silhouette width of the silhouette plot, defined as the average of the *s(i)* over all objects *i* in the dataset.

In general *pam* is proposed to run several times, each time with a different *k*, and to compare the resulting *silhouette plots*. The user can then select that value of *k* yielding the highest average silhouette width, over all *k*, which is called the silhouette coefficient. Experience has led to the subjective interpretation of the silhouette coefficient (*SC*). This interpretation does not depend on the number of objects (*Table 2*).

Table 2 Interpretation of the silhouette coefficient (*SC*) for partitioning methods

SC	Proposed interpretation
0.71-1.00	A strong structure has been found
0.51-0.70	A reasonable structure has been found
0.26-0.50	The structure is weak and could be artificial, try additional methods
≤ 0.25	No substantial structure has been found

Fuzzy Analysis: function „fanny”

The functions *pam* is a *„crisp”* clustering method. This means that each object of the dataset is assigned to exactly one cluster. For instance, an object lying between two clusters will be assigned to one of them. However, a fuzzy method spreads each object over the various clusters. For each object *i* and each cluster *v* there will be a membership u_{iv} which indicates how strongly object *i* belongs to cluster *v*. Memberships have to satisfy the following conditions:

$$u_{iv} \ge 0 \text{ for all } i = 1, \ldots, n \text{ and all } v = 1, \ldots, k.$$

$$\sum_{v=1}^{k} u_{iv} = 1 = 100\ \% \text{ for all } i = 1, \ldots, n.$$

We will focus on the method *fanny* (*Kaufman and Rousseeuw*, 1990), where the memberships $u_{i\nu}$ are defined through minimization of the objective function:

$$\sum_{v=1}^{k} \frac{\sum_{i,j=1}^{n} u_{iv}^{2} u_{jv}^{2} d(i,j)}{2\sum_{j=1}^{n} u_{jv}^{2}} . \tag{8}$$

In this expression, the dissimilarities *d(i; j)* are known and the memberships $u_{i\nu}$ are unknown. The minimization is carried out numerically by means of an iterative algorithm, taking into account the side constraints on memberships by means of Lagrange multipliers.

Compared to other fuzzy clustering methods, *fanny* has the advantage that it can handle dissimilarity data, since Eq. 8 uses only inter-object dissimilarities and does not involve any averages of objects (*Rousseeuw*, 1995). Also, *fanny* is rather robust to the assumption of spherical clusters since the *d(i; j)* in (0.5) are not squared.

For any fuzzy clustering, such as the one produced by *fanny*, one can consider the nearest crisp clustering. The latter assigns each object *i* to the cluster *v* in which it has the highest membership $u_{i\nu}$. This crisp clustering can then be represented by a silhouette plot.

3.3.1.2. Hierarchical methods

Agglomerative Nesting: function „agnes"

The function *agnes* is of agglomerative hierarchical type, hence it yields a sequence of clustering. In the first clustering each of the *n* objects forms its own separate cluster. In subsequent steps clusters are merged, until (after *n − 1* steps) only one large cluster remains. Many such methods exist. In *agnes*, the group average method is taken, based on arguments of robustness, monotonicity and consistency. Also four other well-known methods are available in *agnes*, namely single linkage, complete linkage, Ward's method, and weighted average linkage. These five methods can be described in a unified way (*Lance and Williams*, 1966).

The *agglomerative coefficient* (*AC*) (*Rousseeuw*, 1986) measures the clustering structure of the dataset. For each observation *i*, denote by *d(i)* its dissimilarity to the first cluster it is merged with, divided by the dissimilarity of the merger in the last step of the algorithm. *AC* is then defined as the average of all *1 − d(i)*. It can also be seen as the average width (or the percentage filled) of the banner plot. Note that the *AC* tends to increase with the number of objects, unlike the average silhouette width. Because it grows with the number of observations, this measure should not be used to compare datasets of very different sizes.

The AC derived by *agnes* measures the goodness of the analyzed hierarchy.

The hierarchy obtained from *agnes* can be graphically displayed in two ways: by means of a *clustering tree* or by a *banner*.

Agglomerative tree: A tree in which the leaves represent objects. The vertical coordinate of the junction of two branches is the dissimilarity between the corresponding clusters. An agglomerative clustering tree is a rotated version of a dendrogram (*Anderberg*, 1973).

Agglomerative banner: The banner shows the successive mergers from left to right. (Imagine the ragged flag parts at the left, and the flagstaff at the right.) The objects are listed vertically. The merger of two clusters is represented by a horizontal bar which commences at the between-cluster dissimilarity. The banner thus contains the same information as the clustering tree. Note that the agglomerative coefficient (*AC*) defined above can be seen as the average width (the percentage filled) of the banner.

In this study *Ward's* method is used, since it does not depend on extreme values, besides it produces more realistic groupings (*Anderberg*, 1973; *Kalkstein et al.*, 1987; *Hair et al.*, 1998; *Sindosi et al.*, 2003). The database was standardized with the "stand" option of *agnes* before calculating the dissimilarities. In this case the characterization of a distance between two observations *k* and *l* as *"close"* or *"far"* is determined by the square of their Euclidean distance:

$$D_{kl}^2 = \sum_{i=1}^{m} (x_{ki} - x_{li})^2 \tag{9}$$

where x_{ki} is the value of the *i*th factor for the *k*th day and x_{li} is the value of the *i*th factor for the *l*th day.

Monothetic analysis: function „mona"

The function *mona* is a different type of divisive hierarchical method, which operates on a data matrix with binary variables. For each split *mona* uses a single (well-chosen) variable, which is why it is called a monothetic method. Most other hierarchical methods (including *agnes*) are polythetic, i.e. they use all variables simultaneously.

The output of the function *mona* facilitates evaluating a divisive banner, which is defined as follows. Divisive banner: The banner shows the successive mergers from left to right. (Imagine the ragged flag parts at the right, and the flagstaff at the left.) The objects are listed vertically. The merger of two clusters is represented by a horizontal bar which commences at the between-cluster dissimilarity.

Introduce the following notations:

$$E_v = \sum_{\substack{c \in C \\ v(c)=1}} 1 \quad \text{and} \quad N_v = \sum_{\substack{c \in C \\ v(c)=0}} 1 \tag{10}$$

where *C* is an optional group, while v is a contingency table of any v_i and v_j variables:

	N_{v_j}	E_{v_j}
N_{v_i}	$a(v_i, v_j)$	$b(v_i, v_j)$
E_{v_i}	$c(v_i, v_j)$	$d(v_i, v_j)$

Mark

$$con(v_i, v_j) = a(v_i, v_j) d(v_i, v_j) - b(v_i, v_j) c(v_i, v_j), \tag{11}$$

relation of the variables v_i and v_j. The total association of a variable v is defined by the following formula:

$$\sum_{i=1}^{|V|} con(v, v_i), \tag{12}$$

where V is the set of variables. The variable used for splitting a cluster is the variable with the maximal total association to the other variables, according to the observations in the cluster to be splitted.

A cluster is divided into one cluster with all observations having value 1 for that variable, and another cluster with all observations having value 0 for that variable.

The clustering hierarchy constructed by *mona* can also be represented by means of a divisive banner. The length of a bar is now given by the number of divisive steps needed to make that split. Inside the bar, the variable is listed which was responsible for the split. A bar continuing to the right margin indicates a cluster that cannot be split.

All statistical computations were performed with SPSS (version 9.0) and R softwares.

Acknowledgements – The authors thank B. Vaskövi and B. László for handing over Hungarian national immission data for the period October 1, 2000 – March 31, 2001. This paper was supported by the Bolyai Research Scholarship of the Hungarian Academy of Sciences (BO/00519/07).

REFERENCES

Anderberg, M.R., 1973: *Cluster Analysis for Applications.* Academic Press, New York. 353 p.

Bartzokas, A. and Metaxas, D.A., 1993: Covariability and climatic changes of the lower troposphere temperatures over the Northern Hemisphere. *Nouvo Cimento della Societa Italiana di Fisica C Geophysics and Space Physics 16C*, 359-373.

Bartzokas, A. and Metaxas, D.A., 1995: Factor Analysis of Some Climatological Elements in Athens, 1931-1992: Covariability and Climatic Change. *Theor. and Appl. Climato. 52*, 195-205.

Cutter, S.L., 1992: Green Cities. Ranking major cities by environmental quality reveals some surprises. In: *Hammond, A.* (ed.): *Environmental Almanac*. World Resources Institute – Houghton Mifflin Company, Boston. 169-186.

Gould, S.J., 1996: *Full house: The spread of excellence from Plato to Darwin.* Harmony, New York. 38-42.

Hair, J.F., Anderson, R.E., Tatham, R.L. and Black, W.C., 1998: *Multivariate data analysis.* 5th ed. Prentice Hall, New Jersey. 730 p.

HCSO, 2000a: Budapest Statisztikai Évkönyve, 2000. [*Statistical Year Book of Budapest, 2000.* (in Hungarian)] HCSO, Budapest.

HCSO, 2000b: *Megyei Statisztikai Évkönyvek, 2000. [Statistical Year Books of the Hungarian Counties, 2000.* (in Hungarian)] HCSO, Budapest.

Jolliffe, I.T., 1990: Principal component analysis: A beginner's guide – I. Introduction and application. *Weather 45*, 375-382.

Jolliffe, I.T., 1993: Principal component analysis: A beginner's guide – II. Pitfalls, myths and extensions. *Weather 48*, 246-253.

Kalkstein, L.S., Tan, G. and Skindlov, J.A., 1987: An evaluation of three clustering procedures for use in synoptic classification. *Journal of Climate and Applied Meteorology 26*, 717-730.

Kaufman, L. and Rousseeuw, P.J., 1990: *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley-Interscience, New York. (Series in Applied Probability and Statistics). 342 p. ISBN 0-471-87876-6

Kerényi, A., 1995: *Általános körenyezetvédelem. Globális gondok, lehetséges megoldások.* [*General Environmental Protection. Global concerns, possible solutions.* (in Hungarian)] Mozaik Educational Studio, Szeged. 383 p. ISBN 963 8024 75 5

Lance, G.N. and Williams, W.T., 1966: A general theory of classificatory sorting strategies: 1. Hierarchical systems. *Computer Journal 11*, 195.

Rousseeuw, P.J., 1986: A Visual Display for Hierarchical Classification. In: *Diday, E., Escoufier, Y., Lebart, L., Pages, J., Schektman, Y. and Tomassone, R.* (eds.): *Data Analysis and Informatics 4*. North-Holland. 743–748.

Rousseeuw, P.J., 1995: Fuzzy Clustering at the Intersection. *Technometrics 37*, 283–286.

Sindosi, O.A., Katsoulis, B.D. and Bartzokas, A., 2003: An objective definition of air mass types affecting Athens, Greece; the corresponding atmospheric pressure patterns and air pollution levels. *Environmental Technology 24*, 947-962.

Vaskövi, B., 2000: Országos levegőminőségi (immissziós) adatok 2000. április- szeptember, nem fűtési félév. [National air quality (immission) data 2000 April - September, non-heating half-year. (in Hungarian)] *Egészségtudomány 44(4)*, 366-377.