Environmental Modelling & Software 57 (2014) 52–59

Contents lists available at ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

Classification into homogeneous groups using combined cluster and discriminant analysis

József Kovács [a], Solt Kovács [1], Norbert Magyar [a], Péter Tanos [b], István Gábor Hatvani [a,*], Angéla Anda [b]

[a] Eötvös Loránd University, Department of Physical and Applied Geology, H-1117 Budapest, Pázmány P. stny. 1/C, Hungary
[b] Pannon University, Department of Sanitary and Environmental Engineering, H-8360 Keszthely, Festetics u. 7, Hungary

ARTICLE INFO

Article history:
Received 5 July 2013
Received in revised form
20 December 2013
Accepted 6 January 2014
Available online 6 February 2014

Keywords:
Classification into homogeneous groups
Combined cluster and discriminant analysis (CCDA)
Monitoring network
R package

ABSTRACT

The classification of observations into groups is a general procedure in modern research. However, when searching for homogeneous groups the difficulty of deciding whether further division of a classification is necessary or not to obtain the desired homogeneous groups arises. The presented method, Combined cluster and discriminant analysis (CCDA), aims to facilitate this decision.

CCDA consists of three main steps: (I) a basic grouping procedure; (II) a core cycle where the goodness of preconceived and random classifications is determined; and (III) an evaluation step where a decision has to be made regarding division into sub-groups. These steps of the proposed method were implemented in *R* in a package, under the name of ccda.

To present the applicability of the method, a case study on the water quality samples of Neusiedler See is presented, in which CCDA classified the 33 original sampling locations into 17 homogeneous groups, which could provide a starting point for a later recalibration of the lake's monitoring network.

Software availability

Name of software: Combined cluster and discriminant analysis (CCDA)
Developers: Solt Kovács, József, Kovács, Péter Tanos
Contact address: József, Kovács, Eötvös Loránd University, Department of Physical and Applied Geology, H-1117 Budapest, Pázmány P. stny. 1/C., Hungary. Email: ccda@caesar.elte.hu
Availability and Online Documentation: Free download with description at: http://cran.r-project.org/web/packages/ccda/
Year first available: 2014
Hardware and software required: PC or Mac with any operating system, which is compatible with *R* (freely available at: http://cran.r-project.org/)
Programming language: R language
Program size: 8.69 KB

Abbreviation definition
CCDA: Combined cluster and discriminant analysis
CD: Coded dataset
d: Difference between ratio and the q_{95}
GR: Grouping
GRV: Grouping vector
HCA: Hierarchical cluster analysis
IUCN: International Union for Conservation of Nature
LDA: Linear discriminant analysis
N: The number of sample origins (number of sampling locations)
q_{95}: 95% quantile of the percentages for the random groupings
ratio: Ratio of correctly classified cases for the coded dataset
RCD: Randomly coded dataset
SG: Sub-group
SL: Sampling location
UNESCO: United Nations Educational, Scientific and Cultural Organization

1. Introduction

The classification of different observations into groups is a general procedure in modern research. Be it sampling locations in environmental and earth sciences, species in biology, postal codes for a market research or characteristics of flood retention basins in risk assessment, the question frequently arises, how can one obtain

* Corresponding author. Eötvös Loránd University, Faculty of Science, Department of Physical and Applied Geology, H-1117 Budapest, Pázmány P. stny. 1/C, Hungary. Tel.: +36 70 317 97 58; fax: +36 1 31 91738.
E-mail addresses: kevesolt@geology.elte.hu (J. Kovács), kovacssolt@gmail.com (S. Kovács), magyarnorbert87@gmail.com (N. Magyar), tanospeter@gmail.com (P. Tanos), hatvaniig@gmail.com (I.G. Hatvani), anda-a@georgikon.hu (A. Anda).
[1] Solt Kovács is currently an M.Sc. Student in Mathematics at the ETH Zurich.

1364-8152/$ – see front matter © 2014 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.envsoft.2014.01.010

the largest possible homogeneous groups, for example in order to reduce the number of sampling locations, or to investigate smaller subsystems.

The grouping procedure can be done manually (based on professional experience, intuition etc.) or with the aid of certain methods (e.g. different types of cluster analysis). While facing the difficulty of deciding whether the clusters should be further divided or contracted, the scientist has to make a decision, which is a key element of every clustering procedure (Anderberg, 1973). This issue can be observed in many fields where grouping is mostly obtained from hierarchical cluster analysis (HCA; Day and Edelsbrunner, 1984). As an example, in environmental research the work of Xu et al. (2012) should be mentioned, in which both the spatial and temporal groupings are dealt with in relation to the Zhangweinan River, China. In geology Lee et al. (2012) deal with the classification of forensic soil evidences, while in ecology, McKenna (2003) gives a good example of ameliorating cluster techniques in studying ecological communities. Besides the studies which use already consolidated methodology, there are ones, which approach similar problems, and succeed in solving them with their own, newly developed methodology. For example Rowan et al. (2012) developed a hydromorphological classification and decision-support tool for classifying the physical conditions of lakes and to assess the risk of status deterioration of lakes under the surveillance of the EC Water Framework Directive 2000/60/EC; Yang et al. (2012) discussed a multi-label classification for facilitating the management of flood retention basins. Also concerning the topic of floods the work of Straatsma et al. (2013) should be mentioned who assessed the uncertainty in hydromorphological and ecological model outputs caused by land cover classification errors in the Rhine River floodplain. Besides direct classification, the pre-processing steps are crucial as well. Fernandes et al. (2013) presented a set of uni-dimensional pre-processing methods (imputation, discretization etc.) which are suggested to be adapted to multi-dimensional Bayesian network classifiers in the field of fisheries management.

In the case of every classification, let it be in the field of biology, economy, geology, geography, or basically any other discipline, finding optimal classification is the ultimate aim. Although there are methods (Davies and Bouldin, 1979; Dunn, 1973) indicating optimal classification, these do not assure homogeneity. Hence, unlike in the studies previously mentioned, if one is not only looking for similar, but for homogeneous groups – members/elements of which share equal underlying processes – making the crucial decision on the number of groups needed, is even more difficult.

After a grouping of any kind is obtained, it must be validated. One of the methods generally applied in the validation of cluster results is discriminant analysis, a detailed description of which can be found in the books of Duda et al. (2000) and McLachlan (2004). At first the grouping is extended to the individual observations. These sets are then separated by a linear plane (in the case of linear discriminant analysis; LDA), resulting in the percentage of correctly classified observations.

If groups overlap, then classifying observations correctly into multiple groups is more difficult than correctly assigning them to just a couple. Hence, in general LDA tends to classify more observations correctly if the number of groups is smaller. This makes the validation process even more problematic once looking for homogeneous groups.

Therefore the general question is, on what basis should one choose from possible classifications in order to decide objectively whether the groups obtained are homogeneous or not. A new method, called Combined cluster and discriminant analysis (CCDA), is proposed to make the decision process objective, to assure that homogeneous groups would be obtained. If sampling locations are investigated, then after the homogeneous groups are obtained these can be used to provide a basis, from which:

(1) more detailed and localized information could be extracted (e.g. descriptive statistics, searching for background processes etc.);
(2) a pattern (e.g. a spatial one) indicating homogeneity could be retrieved;
(3) a monitoring network could be recalibrated (potential cost reduction).

In fine the proposed methodology (CCDA) not only intends to find similarly behaving sampling locations (as in the case of Hatvani et al., 2011, 2014a, 2014b), but to find the most detailed differences. Therefore it finds homogeneous groups of sampling locations, members of which share equal, and not only similar underlying processes. Naturally – besides sampling locations – any other observations with known origin could be investigated using CCDA.

2. Materials and methods

2.1. Methodology

In developing CCDA two widely known and applied methods have been combined, HCA and LDA. The former divides data into a hierarchy of clusters, where at the lowest level each item belongs to its own cluster; and at the highest level all items belong to the same cluster; the latter builds a function that separates two labelled classes in an optimal fashion subject to specific statistical metrics (Gross et al., 2010). The output of LDA is eventually a percentage describing the ratio of correctly classified observations by the linear plane.

CCDA consists of three main steps (Fig. 1):

(I) a basic grouping procedure;
(II) a core cycle where the goodness of preconceived and certain random classifications is determined;
(III) an evaluation step based on the results of the core cycle where a decision has to be made whether further division into sub-groups is necessary or not.

Steps (I)–(III) should be repeated for the sub-groups found as long as no further division is recommended at the evaluation step and hence one ends up with homogeneous groups.

Before starting CCDA, one should take care of the necessary data preparation. It should be noted that CCDA provides satisfactory results not only for normally distributed data, but for other types of continuous distributions as well, as long as the violation is caused by skewness rather than outliers. Besides that there should not be missing data.

Let N denote the number of sample origins (in this environmental study, sampling locations). Then, as the first step (I), a basic grouping of the sampling locations $SL_1,\ldots,SL_N$ has to be found. The basic grouping is a set of N different groupings $GR_1,\ldots,GR_N$ that are obtained recursively the following way:

$GR_N = \{\{SL_1\},\ldots,\{SL_N\}\}$ meaning that the N sampling locations form N individual groups;

for i in $N-1,\ldots,1$ grouping GR_i is obtained from grouping GR_{i+1} by merging exactly two groups of GR_{i+1} and keeping all the other groups of GR_{i+1} in GR_i as well. The two merged groups should always be "next to" each other.

This way the grouping GR_i always contains i groups. In particular, in grouping GR_1 every sampling location belongs to the same group: $GR_1 = \{SL_1,\ldots,SL_N\}$. To obtain such a basic grouping it is suggested that HCA be applied to the averages of measured parameters at each sampling location using Ward's method; naturally as long as a reasonable basic grouping is obtained, other methods may be applied as well. If HCA is applied, groupings $GR_1,\ldots GR_N$ are obtained by intersecting the corresponding dendrogram at different linkage distances.

As the second step (II), for every one of the obtained groupings $GR_2,\ldots,GR_N$ the so-called core cycle has to be run. There is no point in running the analysis for the trivial grouping GR_1 where every sampling location belongs to one group. In the core cycle the basic idea is always to compare how well observations belonging to the groups are separated by LDA and whether this separation is significantly better than a random separation, which would indicate non-homogeneous groups in the investigated grouping. The steps in the core cycle for a grouping GR_i, (i in $2,\ldots,N$) of the basic grouping are as follows: