Environ Sci Pollut Res
DOI 10.1007/s11356-015-4664-7

RESEARCH ARTICLE

An open-source software package for multivariate modeling and clustering: applications to air quality management

Xiuquan Wang[1] · **Guohe Huang**[2,3] · **Shan Zhao**[1] · **Junhong Guo**[3]

Received: 18 March 2015 / Accepted: 5 May 2015

Abstract This paper presents an open-source software package, rSCA, which is developed based upon a stepwise cluster analysis method and serves as a statistical tool for modeling the relationships between multiple dependent and independent variables. The rSCA package is efficient in dealing with both continuous and discrete variables, as well as nonlinear relationships between the variables. It divides the sample sets of dependent variables into different subsets (or subclusters) through a series of cutting and merging operations based upon the theory of multivariate analysis of variance (MANOVA). The modeling results are given by a cluster tree, which includes both intermediate and leaf subclusters as well as the flow paths from the root of the tree to each leaf subcluster specified by a series of cutting and merging actions. The rSCA package is a handy and easy-to-use tool and is freely available at http://cran.r-project.org/package=rSCA. By applying the developed package to air quality management in an urban environment, we demonstrate its effectiveness in dealing with the complicated relationships among multiple variables in real-world problems.

Keywords Multivariate modeling · Multivariate clustering · Stepwise cluster analysis · Cluster tree · Air quality management

Responsible editor: Marcus Schulz

Electronic supplementary material The online version of this article (doi:10.1007/s11356-015-4664-7) contains supplementary material, which is available to authorized users.

✉ Guohe Huang
huang@iseis.org

[1] Institute for Energy, Environment and Sustainable Communities, University of Regina, Regina, SK S4S 0A2, Canada

[2] Institute for Energy, Environment and Sustainability Research, UR-NCEPU, University of Regina, Regina, SK S4S 0A2, Canada

[3] Institute for Energy, Environment and Sustainability Research, UR-NCEPU, North China Electric Power University, Beijing 102206, China

Introduction

In many real-world problems, multiple factors and dependent variables are often involved and should be taken into account in the modeling process from a systematic viewpoint. This is especially true in environmental problems, such as air and water pollution, deforestation, soil erosion, biodiversity loss, and global warming (e.g., Cardinale et al. 2012; de Vente et al. 2013; DeFries et al. 2010; Hung et al. 2012; Ring et al. 2012; Wang and Huang 2015; Wang et al. 2014a,d), which often lead to various consequences and are usually caused by a large number of interactive factors associated with human overpopulation, natural resources utilization, economic development, and climatic changes (Jordan et al. 2014; Mellino et al. 2015; Wang et al. 2014b,c; Westing 2013; Xu et al. 2014). Modeling the complex relationships between environmental consequences and their potential incentives is extremely important in management practices, as it can provide scientific bases for environmental prediction, risk assessment, policy analysis, and decision making (Healey et al. 2014; Ma et al. 2014; Zhang et al. 2014).

Such multivariate problems can be converted into a generalized problem that is to model the relationships between

Published online: 14 May 2015

multiple independent and dependent variables. Previously, a number of research efforts have been made to facilitate the development of multivariate relationships through various methods, such as discriminant analysis (e.g., Clemmensen et al. 2011; Ye 2007), neural networks (e.g., Specht 1990; Wasserman 1993), and Bayesian belief networks (e.g., Cooper 1990; Marcot et al. 2001). Among them, the stepwise cluster analysis (SCA) method is one of the most popular methods and has been widely used to handle multivariate modeling problems in environmental management practices. It can effectively deal with continuous and discrete variables, as well as nonlinear relations between the variables. The modeling results of SCA are given by cluster trees, so that a set of forecasting systems, which is flexible to reflect the interactions between multiple independent and dependent variables, can be formed. The SCA method was first introduced by Liu and Wang (1979) to tackle multivariate modeling problems in medical research. Huang (1992) advanced the SCA method and applied it for modeling the relationships between major air pollutants and multiple source factors in an urban environment. Afterwards, a large number of applications based on the SCA method has been reported. For example, Huang et al. (2006) developed a forecasting system for supporting remediation design and process control for NAPL-biodegradation simulation; He et al. (2008) used the SCA to create a set of proxy simulators for quantifying the relationships between operating conditions and benzene levels in a laboratory BTEX system; Qin et al. (2008) developed a dual-phase vacuum extraction process forecasting system for describing the relationships between remediation actions and system responses in an integrated simulation-based system; Sun et al. (2011) employed the SCA method to analyze the complicated interactions between state variables and the carbon/nitrogen (C/N) ratio during food waste composting process; and Wang et al. (2013) applied the SCA method for developing downscaled climate projections over Ontario, Canada.

The wide use of SCA in previous studies (e.g., Bondarenko et al. 1994; Huang et al. 2008; Jiao et al. 2010; Markou et al. 2009; Park et al. 2011; Rúa et al. 1999; Zou et al. 2009) has demonstrated its effectiveness in dealing with the complex interactions between multiple independent and dependent variables. However, to our knowledge, no software for SCA has been developed within a generalized framework such that it can be easily and freely accessed by academia. Therefore, the main objective of this study is to develop an open-source software package that implements the SCA method for general-purpose use in scientific research. The package is named rSCA and can provide a handy and easy-to-use tool for multivariate modeling and analysis. The rSCA package is available from the Comprehensive R Archive Network (CRAN) at http://cran.r-project.org/package=rSCA. This paper is organized as follows: Section 2 reviews the fundamental principles of the SCA method and describes the detailed implementation of the proposed software package; Section 3 presents two applications to air quality management in an urban environment aiming to demonstrate the effectiveness of the rSCA package in handling multivariate relationships in real-world problems; finally, Section 4 presents our summary and conclusions.

Methodology and implementation

The key idea of the SCA method is carrying out a series of cutting (i.e., splitting one set into two) and merging (i.e., joining two sets together) operations to the sample sets of dependent variables while independent variables are used as references in terms of when and how these operations should be taken (Huang 1992; Liu and Wang 1979). The SCA clustering procedure is usually performed as follows (Huang 1992): When no cluster can be cut, mergence of clusters will be performed; when no cluster can be merged with another cluster, cutting action will be carried out; step by step, when all hypotheses of further cutting or mergence are rejected, a cluster tree can then be derived. A typical cluster tree usually consist of intermediate clusters, leaf clusters, as well as a series of cutting and merging rules. If an intermediate cluster can be cut, the cutting rule must be specified to help determine which subcluster a new sample should belong to in the prediction process. A leaf cluster is a sample set that can no longer be cut or merged with others. The mean value of the leaf cluster or an interval bounded by its maximum and minimum values can be used to estimate the predicting results. The prediction is in fact a searching process starting from the top of the tree and ending at a leaf cluster, following the flow path guided by the cutting and merging rules (Wang et al. 2013).

Clustering criterion

The SCA method employs the theory of multivariate analysis of variance (MANOVA) (Cooley and Lohnes 1971; Morrison 1967; Overall and Klett 1983) to help decide whether the difference between two sets of dependent variables is significant or not at a given significance level. In detail, the criterion for cutting or merging are based on the Wilks' lambda statistic (Wilks 1962), which is defined as:

$$\Lambda = \frac{|\mathbf{E}|}{|\mathbf{E}+\mathbf{H}|} \tag{1}$$

where $\mathbf{E}$ and $\mathbf{H}$ are the within-group and between-group sums of squares and cross-products (SSCP) matrices, respectively. Assume two sets of dependent variables $\mathbf{e}$ and $\mathbf{f}$ contain n_e and n_f samples; each individual sample in these two sets can be expressed as: $\mathbf{e}_i=(e_{i1},e_{i2},e_{i3},\ldots,e_{id})$, $i=1,2,3,\ldots,n_e$, or

$\mathbf{f}_j=(f_{j1},f_{j2},f_{j3},\ldots,f_{jd}), j=1,2,3,\ldots,n_f$; here, d is the dimension of **e** and **f**. Then, the **E** and **H** can be given by:

$$\mathbf{E}=\sum_{i=1}^{n_e}(\mathbf{e}_i-\bar{\mathbf{e}})'(\mathbf{e}_i-\bar{\mathbf{e}})+\sum_{j=1}^{n_f}(\mathbf{f}_j-\bar{\mathbf{f}})'(\mathbf{f}_j-\bar{\mathbf{f}}) \tag{2}$$

$$\mathbf{H}=\frac{n_e n_f}{n_e+n_f}(\bar{\mathbf{e}}-\bar{\mathbf{f}})'(\bar{\mathbf{e}}-\bar{\mathbf{f}}) \tag{3}$$

where $\bar{\mathbf{e}}$ is the sample mean of **e**, and $\bar{\mathbf{f}}$ is the sample mean of **f**, respectively. They can be defined as follows:

$$\bar{\mathbf{e}}=\frac{1}{n_e}\sum_{i=1}^{n_e}\mathbf{e}_i \tag{4}$$

$$\bar{\mathbf{f}}=\frac{1}{n_f}\sum_{j=1}^{n_f}\mathbf{f}_j \tag{5}$$

According to Rao's F approximation (Rao 1952), the Wilks' lambda statistic (denoted as Λ) for the above two sets of dependent variables can be transformed to an F statistic, as follows:

$$F\left(d,\ n_e+n_f-d-1\right)=\frac{1-\Lambda}{\Lambda}\cdot\frac{n_e+n_f-d-1}{d} \tag{6}$$

As described in Wilks' likelihood-ratio criterion (Wilks 1962), the smaller the Λ value, the larger the difference between **e** and **f**. Thus, F test can be adopted here to compare the sample means of these two sets. The null hypothesis would be $H_0:\mu_e=\mu_f$ versus the alternative hypothesis $H_1:\mu_e\neq\mu_f$, where μ_e and μ_f are population means of **e** and **f**. Let the significance level be α; the criterion for cutting would be $F\geq F_\alpha$, and H_0 is false, which implies that the difference between these two sets is statistically significant, whereas $F<F_\alpha$ and H_0 is true would be the merging criterion, which indicates that these two sets have no significant difference (Huang 1992).

Clustering procedure

The principle of SCA is to divide a cluster (i.e., a sample set) containing a number of dependent and independent variables into indivisible subclusters, based on a series of cutting and merging operations. Generally, the SCA clustering procedure will start with a cutting operation by which samples in the original cluster will be divided into two groups. After that, merging and cutting operations will be performed step by step until none of the subclusters can be further divided or merged with other subclusters (see the flow chart shown in Fig. 1). Assuming that the original cluster (denoted as **C**) consists of independent variables (denoted as **X**) and dependent variables (denoted as **Y**), as follows:

$$\mathbf{C}=(\mathbf{X},\ \mathbf{Y})=\left[\begin{pmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1n}\\ x_{21} & x_{22} & x_{23} & \cdots & x_{2n}\\ x_{31} & x_{32} & x_{33} & \cdots & x_{3n}\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ x_{m1} & x_{m2} & x_{m3} & \cdots & x_{mn}\end{pmatrix}\begin{pmatrix} y_{11} & y_{12} & y_{13} & \cdots & y_{1d}\\ y_{21} & y_{22} & y_{23} & \cdots & y_{2d}\\ y_{31} & y_{32} & y_{33} & \cdots & y_{3d}\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ y_{m1} & y_{m2} & y_{m3} & \cdots & y_{md}\end{pmatrix}\right] \tag{7}$$

A numeric example for the original cluster (**C**) can be:

$$\mathbf{C}=(\mathbf{X},\ \mathbf{Y})=\left[\begin{pmatrix} 39.9 & 27\\ 9.1 & 8\\ 11.4 & 14\\ 20.5 & 29\\ 27.3 & 26\end{pmatrix}\begin{pmatrix} 29 & 8.9\\ 36 & 9.98\\ 48 & 12.96\\ 50 & 9.84\\ 31 & 8.84\end{pmatrix}\right] \tag{8}$$

where $m=5$, $n=2$, and $d=2$. The clustering procedure of SCA can be summarized as follows:

Step 1. Sequence the original cluster **C** in an ascending order by the kth column of its independent matrix **X**, where $k=1,2,3,\ldots,n$, and n is the total number of independent variables. Denote the sequenced cluster by the kth column as $\mathbf{C}_k$; thus, there would be k sequenced clusters obtained after this step. For the example in Eq. (8), if $k=1$, the sequenced cluster ($\mathbf{C}_1$) would be:

$$\mathbf{C}_1=(\mathbf{X}_1,\ \mathbf{Y}_1)=\left[\begin{pmatrix} 9.1 & 8\\ 11.4 & 14\\ 20.5 & 29\\ 27.3 & 26\\ 39.9 & 27\end{pmatrix}\begin{pmatrix} 36 & 9.98\\ 48 & 12.96\\ 50 & 9.84\\ 31 & 8.84\\ 29 & 8.9\end{pmatrix}\right] \tag{9}$$

Step 2. For each sequenced cluster, $\mathbf{C}_k$, divide its dependent matrix $\mathbf{Y}_k$ into two groups iteratively by the rth row of independent matrix $\mathbf{X}_k$, where $r=1,2,3,\ldots,m-1$, and m is the total number of samples contained in the ordered cluster. Denote the two groups of $\mathbf{Y}_k$ divided by the rth row as $\mathbf{Y}_{kr}^1$ and $\mathbf{Y}_{kr}^2$, Thus, there would be $n\times(m-1)$ possible

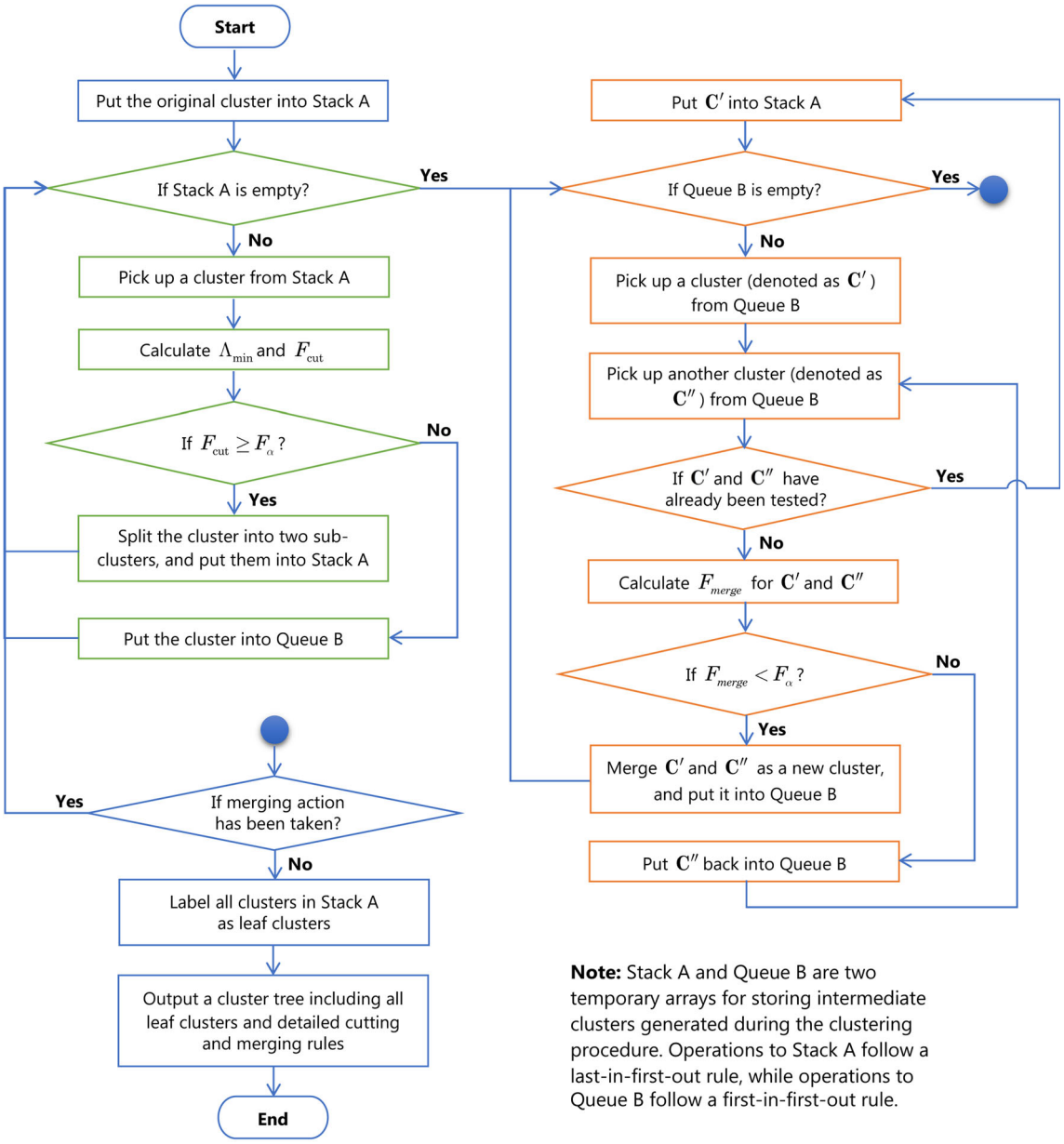

Fig. 1 Flow chart of the SCA clustering procedure

options (labeled as O_{kr}) to divide dependent matrix $\mathbf{Y}_k$. For the example in Eq. (9), if r=2, we have the two divided groups of $\mathbf{C}_1$ as follows:

$$\mathbf{C}_{12}^{1} = (\mathbf{X}_{12}^{1}, \mathbf{Y}_{12}^{1}) = \left[\begin{pmatrix} 9.1 & 8 \\ 11.4 & 14 \end{pmatrix} \begin{pmatrix} 36 & 9.98 \\ 48 & 12.96 \end{pmatrix} \right] \quad (10)$$

$$\mathbf{C}_{12}^{2} = (\mathbf{X}_{12}^{2}, \mathbf{Y}_{12}^{2}) = \left[\begin{pmatrix} 20.5 & 29 \\ 27.3 & 26 \\ 39.9 & 27 \end{pmatrix} \begin{pmatrix} 50 & 9.84 \\ 31 & 8.84 \\ 29 & 8.9 \end{pmatrix} \right] \quad (11)$$

Step 3. For each possible cutting option O_{kr}, calculate its Wilks' lambda statistic Λ_{kr} according to Eq. (1). For example, the Wilks' lambda statistic (i.e., Λ_{12}) for O_{12} can be calculated as follows:

$$\Lambda_{12} = \frac{|\mathbf{E}_{12}|}{|\mathbf{E}_{12} + \mathbf{H}_{12}|} = \frac{\left| \begin{bmatrix} 340.67 & 30.75 \\ 30.75 & 5.07 \end{bmatrix} \right|}{\left| \begin{bmatrix} 340.67 & 30.75 \\ 30.75 & 5.07 \end{bmatrix} + \begin{bmatrix} 34.13 & 14.57 \\ 14.57 & 6.22 \end{bmatrix} \right|} = 0.36 \quad (12)$$

where $\mathbf{E}_{12}$ and $\mathbf{H}_{12}$ are the within- and between-group SSCP matrices of $\mathbf{Y}_{12}^{1}$ and $\mathbf{Y}_{12}^{2}$, respectively.

Step 4. Identify the minimal value among $n \times (m-1)$ Wilks' lambda statistics and denote it as Λ_{min}. Calculate the corresponding F statistic (denoted as F_{cut}) for Λ_{min} according to Eq. (6). Compare F_{cut} to

F_α to decide if the original cluster $\mathbf{C}$ can be cut into two subclusters $\mathbf{C}_{kr}^1$ and $\mathbf{C}_{kr}^2$. If yes, go to Step 5; otherwise, go to Step 6. For the example in Eq. (9), the minimal value, $\Lambda_{\min}$, is reached when k=2 and r=4 (i.e., splitting $\mathbf{C}_1$ by the fourth row while sorted by the second column). Considering a significance level of 0.05 (i.e., α=0.05), we have F_{cut}=54.71>$F_{0.05}$=19. Thus, $\mathbf{C}_1$ can be cut into two subclusters as follows:

$$\mathbf{C}_{24}^1 = \left(\mathbf{X}_{24}^1,\ \mathbf{Y}_{24}^1\right) = \left[\begin{pmatrix} 9.1 & 8 \\ 11.4 & 14 \\ 27.3 & 26 \\ 39.9 & 27 \end{pmatrix} \begin{pmatrix} 36 & 9.98 \\ 48 & 12.96 \\ 31 & 8.84 \\ 29 & 8.9 \end{pmatrix}\right] \quad (13)$$

$$\mathbf{C}_{24}^2 = \left(\mathbf{X}_{24}^2,\ \mathbf{Y}_{24}^2\right) = [(\,20.5 \quad 29\,)(\,50 \quad 9.84\,)] \quad (14)$$

Step 5. Proceed to split the original cluster $\mathbf{C}$ into two subclusters. For each subcluster, repeat the clustering steps starting from Step 1.

Step 6. Label the original cluster $\mathbf{C}$ as a leaf cluster. Continue to process other subclusters until none of them can be further divided (i.e., all subclusters are labeled as leaf clusters).

Step 7. Pick up any two leaf clusters and calculate the Wilks' lambda statistic and the corresponding F statistic (denoted as F_{merge}). Compare F_{merge} to F_α to decide if they can be merged. If yes, merge them as a new cluster. Repeat this step until no further merging action can be taken.

Step 8. If no merging action is taken in Step 7, go to Step 9; otherwise, repeat the steps starting from Step 1 for all newly merged clusters.

Step 9. Output a cluster tree including all leaf clusters and their corresponding cutting and merging rules.

A schematic example of the cluster tree generated at the end of the clustering procedure is given in Fig. 2. In this example, the cluster tree contains five leaf clusters, which are generated after six cutting and two merging operations. Each leaf cluster may include one or more samples and can be reached by following one or more flow paths. For example, the flow path for leaf cluster 11 in Fig. 2 is unique and can be defined as $x_{*2} \le x_{32} \Rightarrow x_{*3} \le x_{43} \Rightarrow x_{*1} > x_{41}$, but leaf node 15 can be reached through multiple flow paths, including $x_{*2} \le x_{32} \Rightarrow x_{*3} > x_{43} \Rightarrow x_{*1} > x_{21}$, $x_{*2} > x_{32} \Rightarrow x_{*4} \le x_{24} \Rightarrow x_{*1} > x_{21}$, and $x_{*2} > x_{32} \Rightarrow x_{*4} > x_{24} \Rightarrow x_{*3} \le x_{33} \Rightarrow x_{*1} > x_{21}$. In general, all nodes and leaf clusters in a cluster tree are numbered uniquely (i.e., 1, 2, 3, …, n) in the order that they are generated during the clustering procedure. When cutting action is taken to a cluster, there will be two subclusters generated at the same time. In this case, the left subcluster will be numbered first and the right one next. A stack structure, which follows a last-in–first-out rule to operate its elements (as shown in Fig. 1), is used to store newly generated intermediate clusters by the cutting operations. Thus, the right node will always be first processed for cutting test, and the left node will not be dealt with until the right node and its child nodes (if any) cannot be split any more.

Inference and prediction

The cluster tree generated in the clustering procedure of SCA can be regarded as a statistical model to represent the relationships of multiple dependent variables versus multiple independent ones. It thus can be used for inference or prediction of the outcomes of dependent variables, while new information of independent variables is available. The process of inference or prediction is in fact a searching process starting from the top of the cluster tree and ending at a leaf cluster. Take the cluster tree in Fig. 2 as an example, the first step is to compare the value of the second independent variable with the value of x_{32}. If it is greater than x_{32}, then proceed to the right branch of the cluster tree; otherwise, go to the left branch. Similarly, the searching process will continue until a leaf cluster is reached. Then, the samples included in the leaf cluster will be used to estimate the values of dependent variables. A number of options are applicable for calculating the estimates of dependent variables, including mean of the samples, an interval bounded by the maximum and minimum values, mean of the samples together with a radius, which is expressed as half of the distance between the minimum and maximum values, random value between the minimum and maximum values, as well as mean of the samples together with their standard deviation.

Functions of the rSCA package

The rSCA package is implemented with four functions to facilitate multivariate modeling and clustering, as follows:

- rSCA.missing: preliminary checking for missing values in the sample sets of independent and dependent variables. This function can be used to help check if there are any missing values in the original data sets. It prints out general statistics for missing values and their locations in the data matrix. Conducting missing check is helpful for understanding the quality of original data sets, but it is not mandatory. Users may choose to skip this preliminary checking. Note that the rSCA package will automatically remove a sample as long as one element in it is missing.
- rSCA.correlation: preliminary analysis for the correlations between dependent and independent variables. This function aims to analyze the correlations between dependent

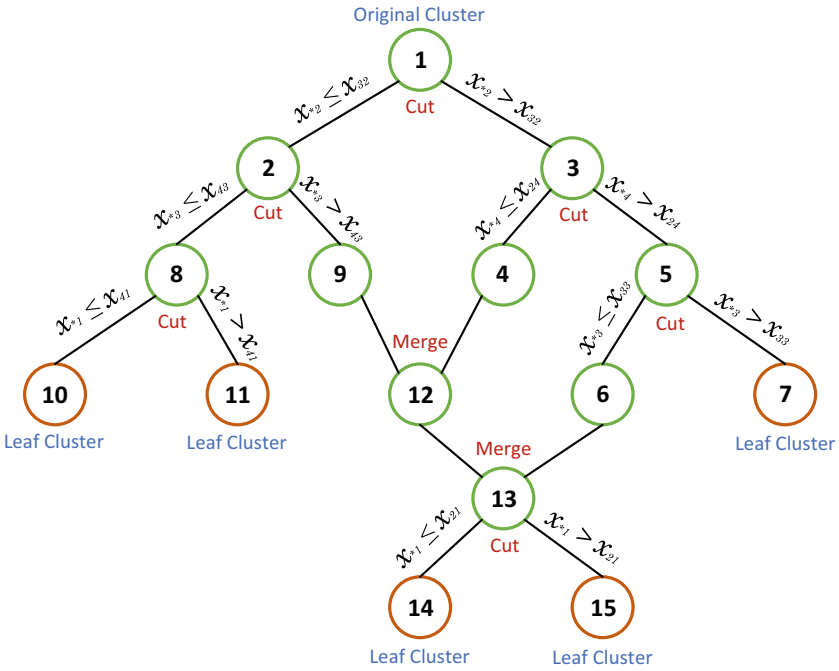

Fig. 2 A schematic cluster tree generated by the SCA clustering procedure

and independent variables. It prints out a table consisting of the correlation coefficient for each pair of dependent and independent variables. Users are recommended to carry out a correlation analysis ahead of the modeling process. Some irrelevant independent variables, which are not correlated with any of dependent variables, can thus be ruled out to reduce the computing time required for the clustering procedure. However, we should note that correlation analysis can be skipped, and the resulting cluster tree will not be affected.

- rSCA.modeling: This function is the key to the rSCA package and serves as a tool for modeling the relationships between dependent and independent variables. The modeling results are represented by a cluster tree. The information for the cluster tree is saved into two plain text files: a tree file with the prefix of "tree" and a map file with the prefix of "map." The tree file stores the structure of the clustered tree, and the map file contains the detailed information of leaf clusters. There two files are usually generated at the current work directory. This function can also run with a debug mode under which more detailed information about the clustering procedure will be recorded. If the debug mode is enabled, a log file with the prefix of "log" will be generated at the current work directory.
- rSCA.inference: multivariate inference or prediction based on a model (i.e., cluster tree) generated through the function of rSCA.modeling. After a cluster tree is obtained, one can employ this function to conduct statistical inference or prediction for new inputs of independent variables. The results will be saved into a text file with the prefix of "rsl" at the current work directory.

More details about these functions are described in the supplementary information (see Text S1).

Applications to air quality management

To demonstrate how the rSCA package can be applied for modeling multivariate relations in real-world problems, we present two applications focused on air quality management in an urban environment. In particular, our case study focus on the city of Xiamen, which is located on the southeast coast of China. The ambient air quality in Xiamen was monitored by a network of stations operated by Xiamen Environmental Monitoring Station (XEMS). The network contained 31 monitoring stations, which were distributed to 31 grid squares (1×1 km^2) covering all the urban districts of Xiamen. Six air pollutants including SO_2, NO_x, O_3, CO, TSP, and dust fall (DF) were monitored by the network. Among them, only three primary pollutants (i.e., SO_2, NO_x, and DF) were screened out in this study through a stepwise discriminant analysis to avoid excessive calculation in the modeling process (Huang 1992). As reported in previous studies (Huang and Sun 1988; Sun 1989), the air pollution in Xiamen was attributed to four major sources: industrial coal consumption, population density, traffic flow, and shopping density. Although there were many other sources more or less affecting the ambient air quality, they were neglected in this study due to these reasons (Tan et al. 2014): (a) they can hardly be quantified in the modeling process (e.g., energy and industrial structures); (b) most of them are associated with four major sources (e.g., the correlation between the total number of vehicles and the gross

Table 1 Data used in the demonstrative applications (Huang 1992)

Station	Major source factors				Primary air pollutants		
	Industrial coal consumption (X_1) (tonne year^{-1})	Population density (X_2) (10^3 km^{-2})	Traffic flow coefficient (X_3)	Shopping density coefficient (X_4)	SO_2 (Y_1) (mg m^{-3})	NO_2 (Y_2) (mg m^{-3})	DF (Y_3) (tonne km^{-2} month^{-1})
1	0.095	0.044	39.9	27	0.02	0.034	10.01
2	0.81	0.058	9.1	8	0.011	0.011	6.92
3	0.101	0.077	11.4	14	0.016	0.018	9.53
4	0.006	0.141	20.5	29	0.022	0.018	5.04
5	0.07	0.281	27.3	26	0.031	0.029	8.9
6	0.481	0.514	30.2	48	0.057	0.036	9.98
7	0.12	0.286	36.4	39	0.04	0.048	12.96
8	0.48	0.199	40.9	27	0.061	0.05	9.84
9	0.112	0.101	29.9	18	0.023	0.031	8.84
10	0.026	0.203	48.1	28	0.025	0.02	4.66
11	0.128	1.235	48.2	61	0.041	0.042	9.02
12	2.681	0.439	51.1	98	0.07	0.029	11.37
13	1.601	0.333	56.1	99	0.077	0.022	11.88
14	1.398	0.455	19.3	103	0.105	0.038	11.06
15	1.256	0.314	14.9	17	0.038	0.027	11.64
16	2.618	0.609	9.1	19	0.058	0.019	8.25
17	1.217	0.88	17.2	73	0.051	0.05	10.01
18	1.411	2.115	19.6	203	0.073	0.038	9.2
19	0.245	6.839	49.2	296	0.123	0.08	9.91
20	0.724	3.06	17.1	192	0.089	0.046	9.37
21	0.019	2.252	29.1	123	0.073	0.039	7.99
22	1.321	5.73	41.1	288	0.139	0.069	13.28
23	0.903	3.078	39	97	0.095	0.048	9.8
24	0.714	1.013	16.7	5	0.034	0.04	8.5
25	0.581	1.398	11.7	57	0.055	0.034	9.21
26	0.08	1.734	10.2	52	0.02	0.05	8.67
27	0.12	1.848	6.6	132	0.07	0.036	8.03
28	0.089	1.357	10.3	148	0.058	0.039	8.01
29	0.112	0.585	19.3	79	0.057	0.031	6.3
30	0.192	0.675	6.9	39	0.05	0.014	7.92
31	0.301	1.937	11.9	6	0.039	0.04	8.08

population); (c) their impacts are less significant (e.g., natural gas and oil consumption) compared with the major sources. In this study, we obtained the data for four major sources and three primary air pollutants (shown in Table 1) from the paper of Huang (1992). Our first application aims to model multivariate relationships among the concentrations of three primary air pollutants, SO_2 (denoted as Y_1), NO_2 (denoted as Y_2), and DF (denoted as Y_3), and four major source factors including industrial coal consumption (denoted as X_1), population density (denoted as X_2), traffic flow coefficient (denoted as X_3), and shopping density coefficient (denoted as X_4) at 31 monitoring stations. The second application will be focused on the clustering of monitoring stations based on their monitored concentrations for the three primary pollutants.

Multivariate modeling for air quality prediction

In this application, we first build a cluster tree to represent the relationships between pollutant concentrations and source factors with the aid of function rSCA.modeling. Then, the cluster tree will be used to predict the corresponding concentrations of three pollutants given new data for source factors. The correlation coefficients between pollutant concentrations and source factors can be obtained through rSCA.correlation (see the supplementary information: Text S2) and are listed in Table 2. The correlation coefficient between Y_2 and X_1 is close to 0, but the effects of X_1 on Y_1 and Y_3 are not ignorable. We therefore should retain X_1 in the modeling process. In practice, correlation analysis may be skipped for small data sets

Table 2 Correlation coefficients of pollutant concentrations versus source factors

Source factors	Pollutant concentrations		
	SO_2 (Y_1)	NO_2 (Y_2)	DF (Y_3)
Industrial coal consumption (X_1)	0.3787	−0.0716	0.4316
Population density (X_2)	0.7523	0.7886	0.2371
Traffic flow coefficient (X_3)	0.3083	0.3189	0.3496
Shopping density coefficient (X_4)	0.8490	0.6722	0.3170

because of their less requirements for computing resources in the clustering procedure, but it is essential for large data sets to check correlation coefficients between independent and dependent variables and drop non- or less-related independent variables. Next, the cluster tree representing the complex relationships between pollutant concentrations and source factors can be obtained using the function of rSCA.modeling (see the supplementary information: Text S2). Here, we set different values for the significance level (i.e., α=0.01, α=0.05, and α=0.1) to test the sensitivity of modeling results. For convenience, we denote the output models corresponding to the three significance levels as A_1, A_2, and A_3. The cluster trees are shown in Figs. 3, 4, and 5, and their output statistics, including total nodes, leaf nodes, and total numbers of cutting and merging actions, as well as the time used for the entire modeling process, are compared in Table 3. Note that the values in each leaf node as illustrated in Figs. 3, 4, and 5 are the means of all samples included by itself. The rSCA software package allows users to choose different options to summarize the samples within each leaf node. This can be implemented by setting the "mapvalue" parameter in the function of rSCA.modeling. A full list of options for "mapvalue" includes mean, max, min, median, interval, radius, variation, and random. More explanations about these options can be found in the supplementary information (see Text S1). Apparently, different values of significance level can lead to distinct modeling results. In other words, the value of α plays an important role on the structure of the cluster tree because an F statistic at a given significance level will be calculated and used as criterion for cutting or merging operations. Model A_3 with α=0.1 generates the most complicated cluster tree among three ones, and it consists of more intermediate and leaf nodes due to more cutting actions than that of the other models. This is because the higher the value of α is, the more relaxed the cutting criterion is. It is easy to conclude that higher value of α will lead to more cutting actions and more leaf nodes. Thus, each model generated by the rSCA package can be calibrated by adjusting the value of α iteratively until the modeled results reach an acceptable level of performance in terms of reproducing the observations.

Next, we use the above three models (i.e., A_1, A_2, and A_3) to predict the concentrations of three pollutants using the data given by Huang (1992). The function of rSCA.inference will be used to do prediction based on the three models (see the supplementary information: Text S2). We then compare the predicted results of the three models to evaluate their individual performance in reproducing the observed concentrations of three pollutants at all monitoring stations. The comparisons are presented with scatter plots of predicted values versus observed ones (see Fig. 6). In addition, the coefficient of determination (denoted as R^2) is used as a quantitative criterion for assessing the model performance. It is interesting to find that model A_2 performs the best among the three models, although its total number of leaf nodes are not the most. The performance of model A_3 is next to that of model A_2 as its R^2 values are very close to those of model A_2, while model A_1 shows relatively poor performance. Thus, model A_2 is

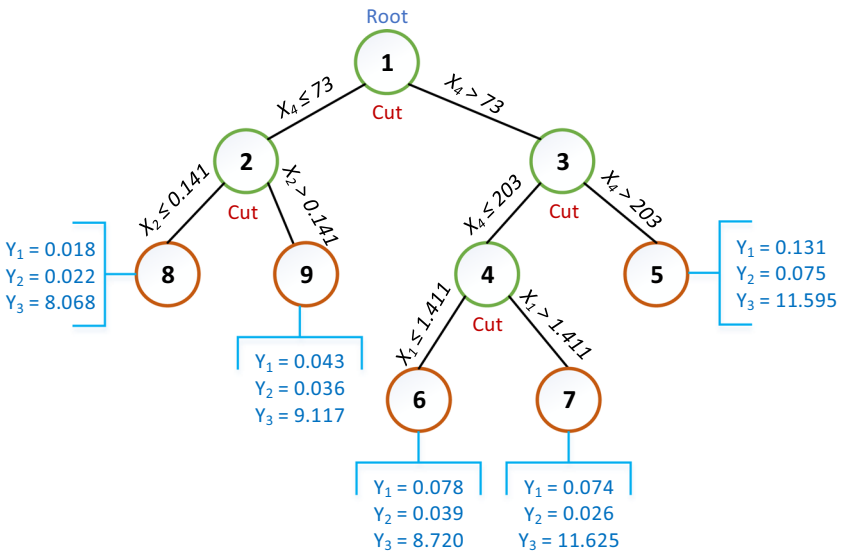

Fig. 3 Cluster tree for model A1 (α=0.01)

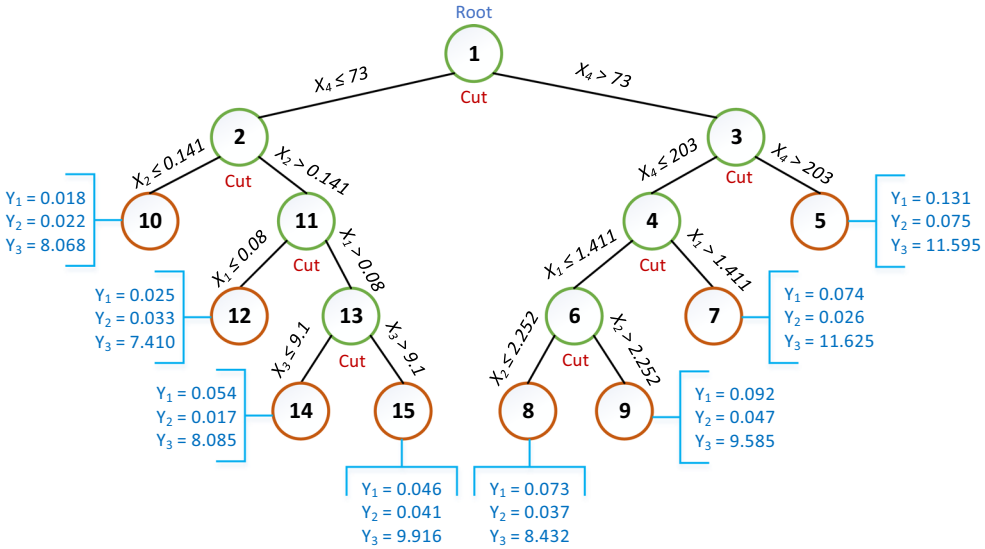

Fig. 4 Cluster tree for model A2 (α=0.05)

recommended for predicting pollutant concentrations, while new data for source factors are available. However, we should note that our recommendation is only based upon a single criterion (i.e., R^2); users may choose to use other metrics in real-world applications to help calibrate or validate their models. In addition, overtraining has been identified as a common issue of statistical models because the parameter values are too intensively optimized on the training set and not optimal in the sense of minimizing the generalization error given by the risk function (Amari et al. 1997; Gardner and Dorling 2000; Jain et al. 2000). Cross-validation is recognized as an effective diagnostic approach to avoid overtraining (Kohavi 1995; Shao 1993) and thus can be used to deal with the similar issue in the course of model development with the rSCA package.

As for the performance of three models in predicting the concentrations of three primary pollutants, we find that the prediction results for DF are relatively poor (with the highest value of R^2 being 0.59) in comparison with those for SO_2 and NO_x (with the R^2 being as high as 0.935 and 0.861, respectively). This may imply that it is reasonable to regard the aforementioned four factors (i.e., industrial coal consumption, population density, traffic flow, and shopping density) as major contributors to SO_2 and NO_x pollutions in the city of Xiamen. However, this may suggest that there might be other

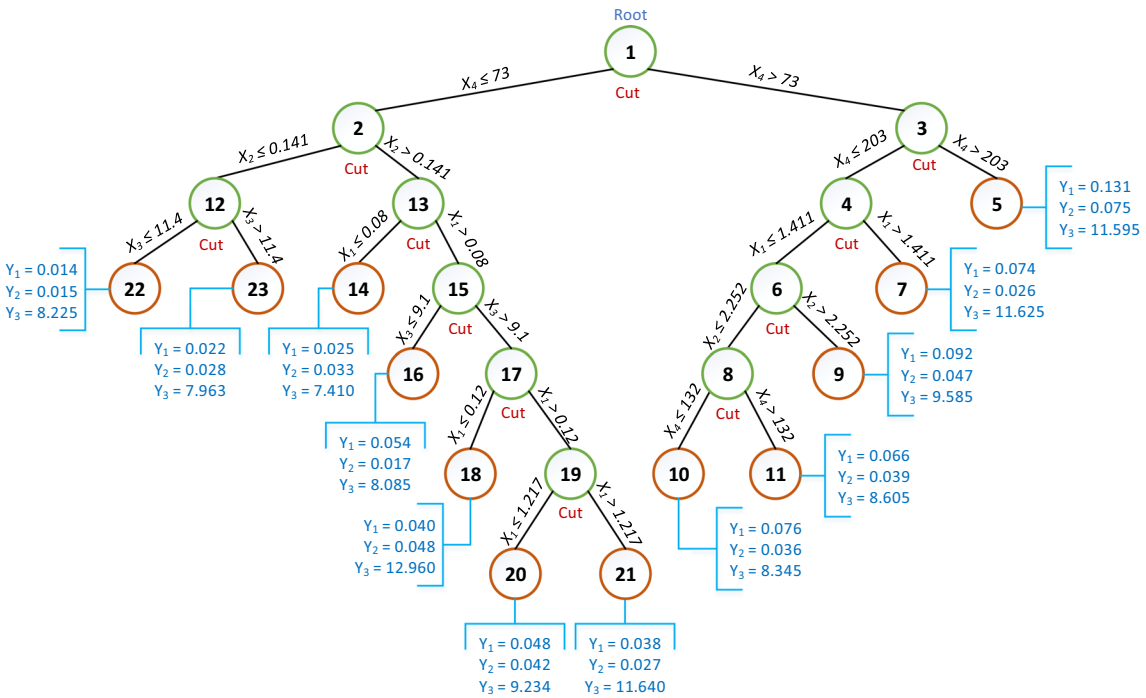

Fig. 5 Cluster tree for model A3 (α=0.1)

Table 3 Comparison of output statistics for three models: A_1, A_2, and A_3

Statistics	Model		
	A_1	A_2	A_3
α	0.01	0.05	0.1
Total nodes	9	15	23
Leaf nodes	5	8	12
Cutting actions	4	7	11
Merging actions	0	0	0
Time used (s)	0.30	0.41	0.39

Note that the above statistics are summarized from the testing results of a personal computer with a 2.9 GHz Intel dual-core processor and 8 GB memory

factors more or less affecting the concentration of DF. These factors should be further identified and quantified to improve the predictability of dust fall in the context of Xiamen.

Multivariate clustering for air quality rating

In this application, we will focus on clustering the 31 monitoring stations into a number of groups based on their measured concentrations for three pollutants (without consideration of the effects of source factors), to support the air quality rating by region in the context of Xiamen. In order to do so, we treat the pollutant concentrations as both independent and dependent variables at the same time. Clustering to the monitoring stations can then be implemented using the function of rSCA.modeling (see the supplementary information: Text S3). The cluster tree containing station information (indicated by station number) is shown in Fig. 7 to help understand the entire clustering process. The 31 monitoring stations are clustered into seven groups: (19, 22), (2, 4, 10), (14, 20, 23), (7, 11, 15, 31), (1, 3, 5, 9, 24, 26), and (6, 8, 16, 17, 25, 28, 29, 30). Note that the significance level for the cluster tree is 0.05; users may choose different values of significance levels for their own applications to obtain appropriate clustering results.

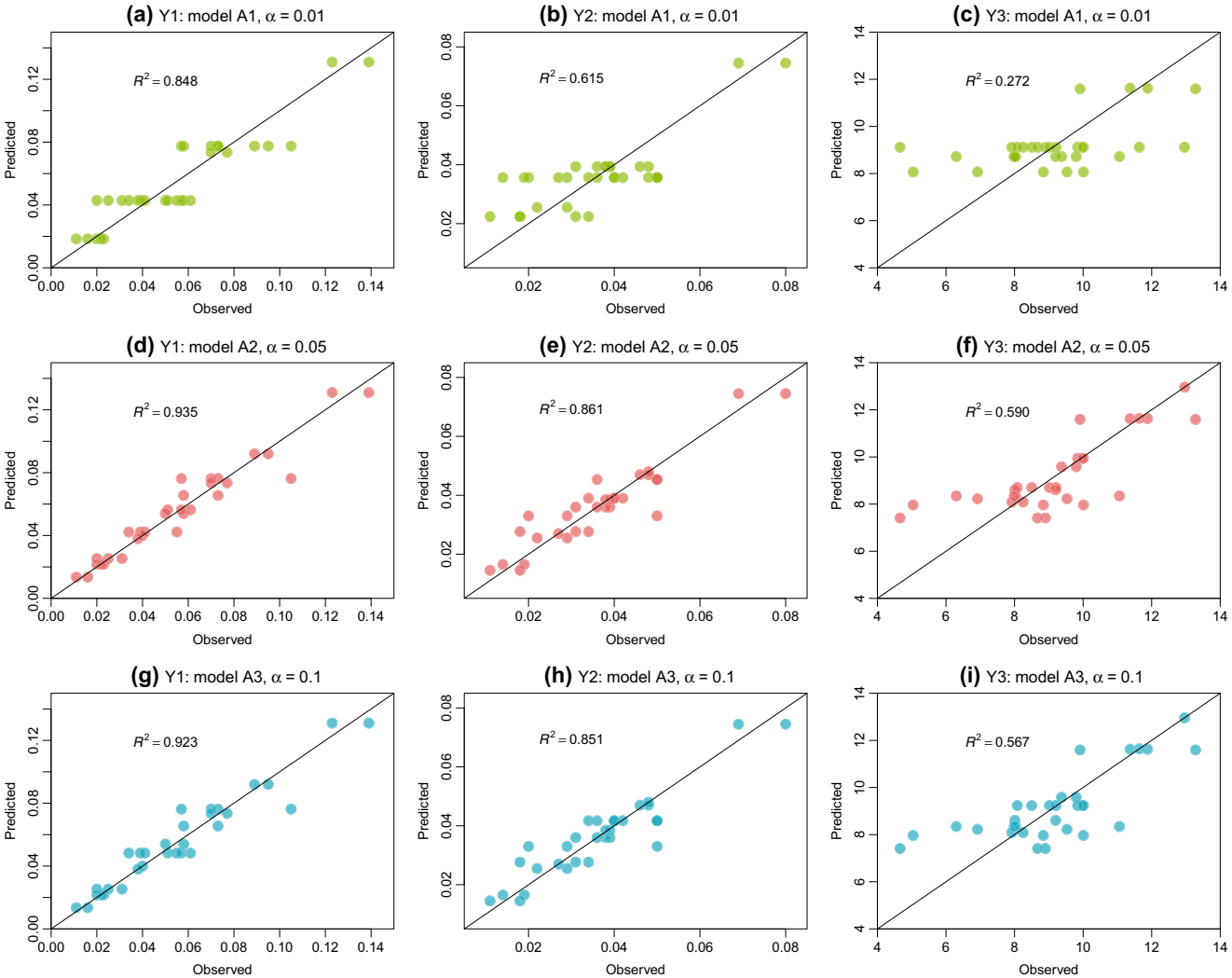

Fig. 6 Comparison between the predicted and observed concentrations of three major pollutants

To compare the clustering performance of rSCA with that of other methods, we use the hierarchical clustering analysis (HCA) function (i.e., "hclust") of R software to cluster the 31 monitoring stations. The "hclust" function provides different options for its agglomeration method to allow for different clustering results. Each agglomeration method is specified by an unambiguous abbreviation. There are eight agglomeration options in total: "ward.D," "ward.D2," "single," "complete," "average," "mcquitty," "median," and "centroid;" users can refer to the R User Manual (available at: http://cran.r-project.org/doc/manuals/R-intro.pdf) for more details about these options. Here, we consider all the agglomeration options in the HCA clustering. The results are presented in Fig. 8. Since there are no best solutions for the problem of determining the number of clusters to extract, we consider seven clusters for every dendrogram (indicated by red rectangles) such that the HCA results are comparable to that of rSCA. For example, the seven clusters for the option of "ward.D" are (2, 29), (4, 10), (7, 22), (12, 13, 14, 15), (1, 6, 8, 17, 19, 23), (16, 21, 27, 28, 30, 31), and (3, 5, 9, 11, 18, 20, 24, 25, 26). Options of "ward.D2," "average," and "centroid" generate the same clusters as option "ward.D" does; options of "complete" and "mcquitty" produce the same results, but they are different from those of "ward.D;" the remaining two options (i.e., "single" and "median") output different results from each other, and their results are also different from others. Not surprisingly, the clusters generated by the rSCA are distinct from all of the HCA results. This is mainly because the rSCA follows a top-to-bottom clustering approach (i.e., putting all objects together as only one cluster at the first beginning and then performing cutting and merging tests iteratively until there is no cutting or merging action to be taken) totally different from the HCA, which applies a bottom-to-top approach (i.e., each object is initially assigned to its own cluster and then proceeding to join the most similar clusters iteratively until there is just a single cluster). While there are no generalized criterion for deciding which method is the best, we only conclude from the comparison that the rSCA can be used as an alternative clustering method because it indeed generates different results. Besides, our comparison can provide important insights into the differences between rSCA and HCA.

Summary and conclusions

This paper implements the SCA method as an open-source software package, which is named rSCA, and can be used as a statistical tool for multivariate modeling and clustering. The modeling results of rSCA are presented as a cluster tree, which includes both intermediate and leaf subclusters as well as the flow paths from the root of the tree to each leaf subclusters specified by a series of cutting and merging operations. The inference or prediction process is straightforward and in fact a searching process starting from the root of the cluster tree and ending at a leaf node, following the flow path guided by a series of cutting and merging rules. The main advantage of rSCA is that it can handle both continuous and discrete variables, as well as nonlinear relationships between multiple variables. This is due to that it build the complex relationships as a cluster tree, and thus, no assumption on the mathematical function is required. The developed software package is then

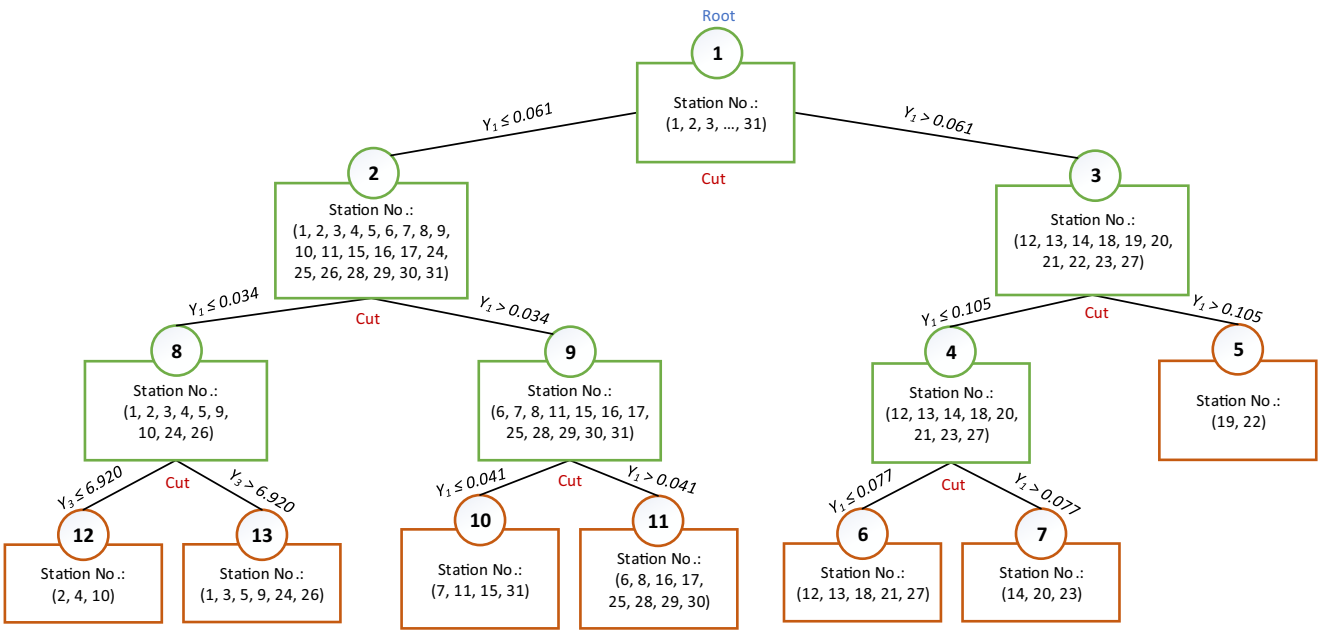

Fig. 7 Clustering process of monitoring stations

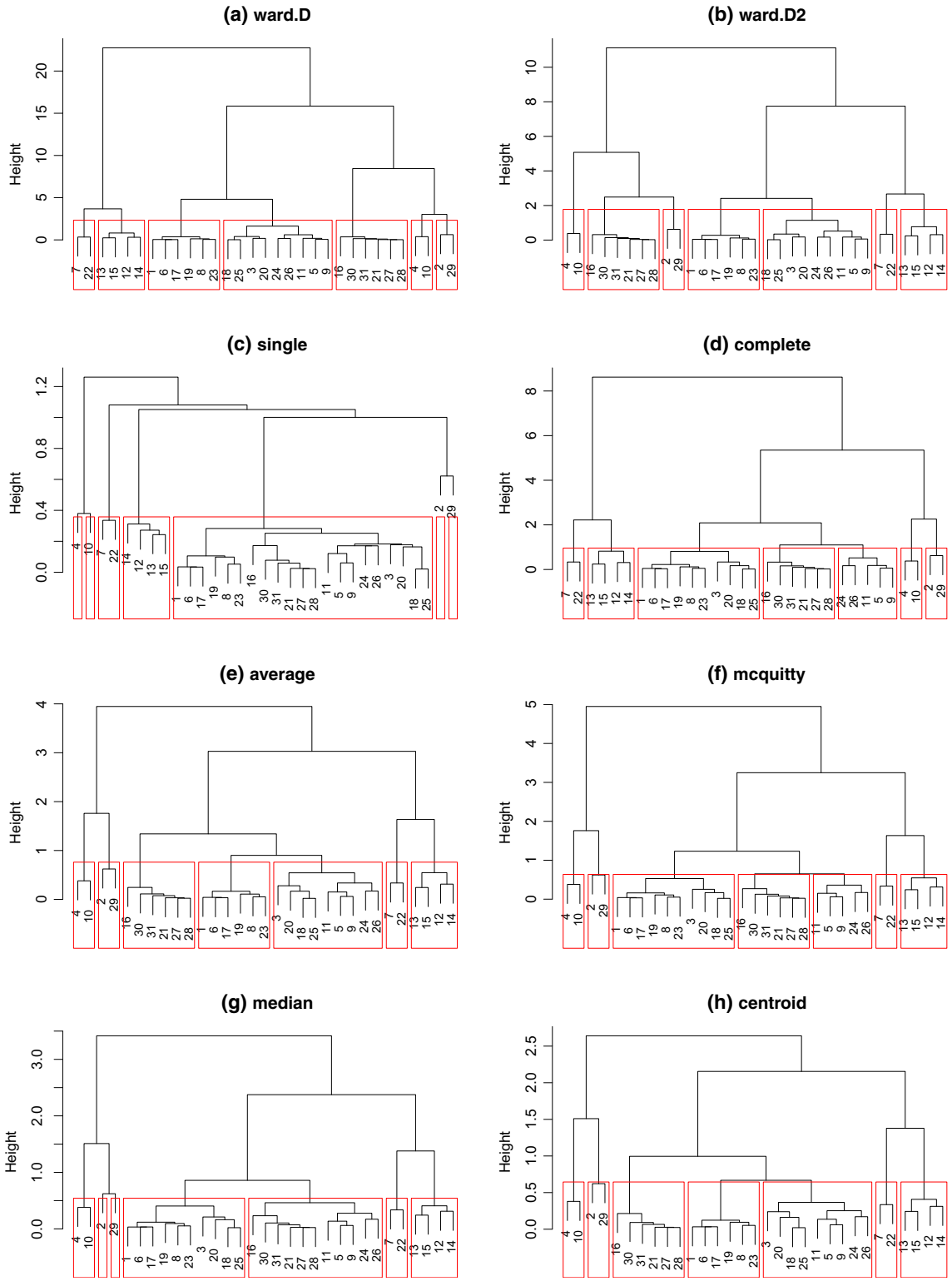

Fig. 8 Results for hierarchical clustering analysis

applied to air quality management in an urban environment, aiming to demonstrate its effectiveness for multivariate modeling and clustering. The results show that the rSCA package can provide both researchers and practical users with an effective tool for handling multivariate relationships in real-world applications.

However, we should note that the two applications presented in this paper depend on a number of limitations or assumptions

that may impose some caveats on applying the rSCA package for real-world cases. First, we only consider four major emission sources to build the air quality prediction model, assuming that air pollution issues in Xiamen can be attributed to four major sources, i.e., industrial coal consumption, population density, traffic flow, and shopping density, while other sources are negligible in comparison with the major ones; in addition, we assume that the local meteorological and geographical conditions do not change significantly with time (Tan et al. 2014). These assumptions seem to be reasonable for such a specific case study because they are based on a preliminary evaluation of the contribution of each source (Huang and Sun 1988; Sun 1989). Nevertheless, all potential sources should be taken into account in real-world applications from a systematic perspective such that the resulting model can reflect the interactions among all sources as well as their causal effects on the ambient air quality. Moreover, such a systematic perspective is very important and indispensable while extending the modeling efforts with rSCA to other environmental prediction problems. Second, only three primary pollutants (i.e., SO_2, NO_x, and DF) are screened out from six monitored pollutants through a stepwise discriminant analysis to avoid excessive calculation in the modeling process (Huang 1992). However, this does not mean that the rSCA model is unable to cover other three pollutants. In fact, the rSCA method itself is capable of taking all primary and secondary pollutants into account at once as long as the concentration data for each pollutant is available and of good quality. In other words, air quality management problems should be tackled on a case-by-case basis as each real-world application may be subject to inevitable assumptions due to limited data availability, poor data quality, or other policy consideration. Finally, there are indeed a few statistical models dedicated for air quality studies, such as chemical mass balance (CMB), positive matrix factorization (PMF), and Unmix. These models are usually known as source receptor ones because they use linear combinations of input sources to identify the contribution of different source sites on a receptor site (Clarkea et al. 2012). The rSCA model developed in our case study is different from these models in three aspects: (1) It aims to build to the statistical relationships between major sources (i.e., industrial and traffic emissions) or factors (i.e., population and shopping density) and the concentrations of primary pollutants at multiple monitoring stations; (2) modeling results are given by a cluster tree without any linear or functional assumptions; and (3) the rSCA package itself is not restricted to air quality management studies, it can be used as an effective tool for tackling multivariate modeling and clustering issues in many other environmental problems.

Acknowledgments This research was supported by the Program for Innovative Research Team in University (IRT1127), the 111 Project (B14008), and the Natural Science and Engineering Research Council of Canada.

References

Amari S-I, Murata N, Muller K-R, Finke M, Yang HH (1997) Asymptotic statistical theory of overtraining and cross-validation. IEEE Trans Neural Netw 8(5):985–996

Bondarenko I, Van Malderen H, Treiger B, Van Espen P, Van Grieken R (1994) Hierarchical cluster analysis with stopping rules built on Akaike's information criterion for aerosol particle classification based on electron probe X-ray microanalysis. Chemom Intell Lab Syst 22(1):87–95

Cardinale BJ, Duffy JE, Gonzalez A, Hooper DU, Perrings C, Venail P, Narwani A, Mace GM, Tilman D, Wardle DA, Kinzig AP, Daily GC, Loreau M, Grace JB, Larigauderie A, Srivastava DS, Naeem S (2012) Biodiversity loss and its impact on humanity. Nature 486(7401):59–67

Clarkea K, Romainb AC, Locogea N, Redona N (2012) Application of chemical mass balance methodology to identify the different sources responsible for the olfactory annoyance at a receptor-site. Chem Eng. 30

Clemmensen L, Hastie T, Witten D, Ersbøll B (2011) Sparse discriminant analysis. Technometrics 53(4)

Cooley WW, Lohnes PR (1971) Multivariate data analysis. J. Wiley

Cooper GF (1990) The computational complexity of probabilistic inference using Bayesian belief networks. Artif Intell 42(2):393–405

de Vente J, Poesen J, Verstraeten G, Govers G, Vanmaercke M, Van Rompaey A, Arabkhedri M, Boix-Fayos C (2013) Predicting soil erosion and sediment yield at regional scales: where do we stand? Earth Sci Rev 127:16–29

DeFries RS, Rudel T, Uriarte M, Hansen M (2010) Deforestation driven by urban population growth and agricultural trade in the twenty-first century. Nat Geosci 3(3):178–181

Gardner M, Dorling S (2000) Statistical surface ozone models: an improved methodology to account for non-linear behaviour. Atmos Environ 34(1):21–34

He L, Huang G-H, Lu H-W, Zeng G-M (2008) Optimization of surfactant-enhanced aquifer remediation for a laboratory BTEX system under parameter uncertainty. Environ Sci Technol 42(6):2009–2014

Healey NC, Oberbauer SF, Ahrends HE, Dierick D, Welker JM, Leffler AJ, Hollister RD, Vargas SA, Tweedie CE (2014) A mobile instrumented sensor platform for long-term terrestrial ecosystem analysis: an example application in an arctic tundra ecosystem. J Environ Inform 24(1):1–10

Huang G (1992) A stepwise cluster analysis method for predicting air quality in an urban environment. Atmos Environ Part B 26(3):349–357

Huang G, Huang Y, Wang G, Xiao H (2006) Development of a forecasting system for supporting remediation design and process control based on NAPL-biodegradation simulation and stepwise-cluster analysis. Water Resour Res 42(6)

Huang G, Sun S (1988) Environmental quality reports of Xiamen Special Economic Zone. Xiamen Environmental Protection Bureau, Xiamen

Huang Y, Wang G, Huang G, Xiao H, Chakma A (2008) IPCS: an integrated process control system for enhanced in-situ bioremediation. Environ Pollut 151(3):460–469

Hung YT, Wang LK, Shammas NK (2012) Handbook of environment and waste management: air and water pollution control. World Scientific

Jain AK, Duin RPW, Mao J (2000) Statistical pattern recognition: a review. IEEE Trans Pattern Anal Mach Intell 22(1):4–37

Jiao S, Zeng G-M, He L, Huang G-H, Lu H-W, Gao Q (2010) Prediction of dust fall concentrations in urban atmospheric environment through support vector regression. J Cent S Univ Technol 17:307–315

Jordan YC, Ghulam A, Chu ML (2014) Assessing the impacts of future urban development patterns and climate changes on total suspended sediment loading in surface waters using geoinformatics. J Environ Inform 24(2):65–79

Kohavi R (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. IJCAI, pp. 1137–1145

Liu Y, Wang Y (1979) Application of stepwise cluster analysis in medical research. Sci Sinica 22(9):1082–1094

Ma ZZ, Wang ZJ, Xia T, Gippel CJ, Speed R (2014) Hydrograph-based hydrologic alteration assessment and its application to the yellow river. J Environ Inform 23(1):1–13

Marcot BG, Holthausen RS, Raphael MG, Rowland MM, Wisdom MJ (2001) Using Bayesian belief networks to evaluate fish and wildlife population viability under land management alternatives from an environmental impact statement. For Ecol Manag 153(1):29–42

Markou MT, Bartzokas A, Kambezidis HD (2009) Daylight climatology in Athens, Greece: types of diurnal variation of illuminance levels. Int J Climatol 29(14):2137–2145

Mellino S, Buonocore E, Ulgiati S (2015) The worth of land use: a GIS-emergy evaluation of natural and human-made capital. Sci Total Environ 506:137–148

Morrison DF (1967) Multivariate statistical methods. McGraw-Hill Book

Overall JE, Klett CJ (1983) Applied multivariate analysis. RE Krieger Publishing Company

Park Y-C, Jeong J-M, Eom S-I, Jeong U-P (2011) Optimal management design of a pump and treat system at the industrial complex in Wonju, Korea. Geosci J 15(2):207–223

Qin X, Huang G, Zeng G, Chakma A (2008) Simulation-based optimization of dual-phase vacuum extraction to remove nonaqueous phase liquids in subsurface. Water Resour Res 44(4)

Rao CR (1952) Advanced statistical methods in biometric research

Ring MJ, Lindner D, Cross EF, Schlesinger ME (2012) Causes of the global warming observed since the 19th century. Atmos Climate Sci 2(04):401

Rúa A, Bourhim S, Marín E, Hernández E (1999) Characterising SO_2 and sulphate patterns in Europe: a cluster analysis. Toxicol Environ Chem 71(1–2):21–32

Shao J (1993) Linear model selection by cross-validation. J Am Stat Assoc 88(422):486–494

Specht DF (1990) Probabilistic neural networks. Neural Netw 3(1):109–118

Sun S (1989) Principal component analysis of air pollutant sources in Xiamen, China. China Environ Sci 10:23–41

Sun W, Huang GH, Zeng G, Qin X, Yu H (2011) Quantitative effects of composting state variables on C/N ratio through GA-aided multivariate analysis. Sci Total Environ 409(7):1243–1254

Tan Q, Wei Y, Wang M, Liu Y (2014) A cluster multivariate statistical method for environmental quality management. Eng Appl Artif Intell 32:1–9

Wang X, Huang G (2015) Impacts assessment of air emissions from point sources in Saskatchewan, Canada—a spatial analysis approach. Environ Prog Sustainable Energy 34(1):304–313

Wang X, Huang G, Lin Q, Liu J (2014a) High-resolution probabilistic projections of temperature changes over Ontario, Canada. J Climate 27(14):5259–5284

Wang X, Huang G, Lin Q, Nie X, Cheng G, Fan Y, Li Z, Yao Y, Suo M (2013) A stepwise cluster analysis approach for downscaled climate projection—a Canadian case study. Environ Model Softw 49:141–151

Wang X, Huang G, Lin Q, Nie X, Liu J (2014b) High-resolution temperature and precipitation projections over Ontario, Canada: a coupled dynamical-statistical approach. Q J R Meteorol Soc

Wang X, Huang G, Liu J (2014c) Projected increases in intensity and frequency of rainfall extremes through a regional climate modeling approach. J Geophys Res Atmos 119(23):13271–13286

Wang X, Huang G, Liu J, (2014d) Projected increases in near-surface air temperature over Ontario, Canada: a regional climate modeling approach. Clim Dyn 1–13

Wasserman PD (1993) Advanced methods in neural computing. John Wiley & Sons, Inc

Westing AH (2013) Population: perhaps the basic issue, from environmental to comprehensive security. Springer, pp. 133–145

Wilks S (1962) Mathematics statistics. John Wiley and Sons, New York

Xu Y, Huang GH, Cheng GH, Liu Y, Li YF (2014) A two-stage fuzzy chance-constrained model for solid waste allocation planning. J Environ Inform 24(2):101–110

Ye J (2007) Least squares linear discriminant analysis, Proceedings of the 24th international conference on Machine learning. ACM, pp. 1087–1093

Zhang N, Li YP, Huang WW, Liu J (2014) An inexact two-stage water quality management model for supporting sustainable development in a rural system. J Environ Inform 24(1):52–64

Zou Y, Huang GH, Nie X (2009) Filtered stepwise clustering method for predicting fate of contaminants in groundwater remediation systems: a case study in western Canada. Water Air Soil Pollut 199(1–4):389–405