Computational Statistics and Data Analysis 79 (2014) 14–29

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Exploratory data analysis of interval-valued symbolic data with matrix visualization

Chiun-How Kao [a,b], Junji Nakano [c], Sheau-Hue Shieh [d], Yin-Jing Tien [b], Han-Ming Wu [e], Chuan-kai Yang [a], Chun-houh Chen [b,*]

[a] Department of Information Management, National Taiwan University of Science and Technology, Taipei 10607, Taiwan
[b] Institute of Statistical Science, Academia Sinica, Taipei 11529, Taiwan
[c] The Institute of Statistical Mathematics, Tokyo 106-8569, Japan
[d] Center for Teacher Education, National Taipei University, New Taipei City 23741, Taiwan
[e] Department of Mathematics, Tamkang University, New Taipei City 25137, Taiwan

ARTICLE INFO

Article history:
Received 17 June 2012
Received in revised form 28 November 2013
Accepted 15 April 2014
Available online 26 April 2014

Keywords:
Symbolic data analysis
Interval-valued data
Matrix visualization
Generalized association plots
Proximity matrix
Exploratory data analysis
EDA

ABSTRACT

Symbolic data analysis (SDA) has gained popularity over the past few years because of its potential for handling data having a dependent and hierarchical nature. Amongst many methods for analyzing symbolic data, exploratory data analysis (EDA: Tukey, 1977) with graphical presentation is an important one. Recent developments of graphical and visualization tools for SDA include zoom star, closed shapes, and parallel-coordinate-plots. Other studies project high dimensional symbolic data into lower dimensional spaces using symbolic data versions of principal component analysis, multidimensional scaling, and self-organizing maps. Most graphical and visualization approaches for exploring symbolic data structure inherit the advantages of their counterparts for conventional (non-symbolic) data, but also their disadvantages. Here we introduce matrix visualization (MV) for visualizing and clustering symbolic data using interval-valued symbolic data as an example; it is by far the most popular symbolic data type in the literature and the most commonly encountered one in practice. Many MV techniques for visualizing and clustering conventional data are converted to symbolic data, and several techniques are newly developed for symbolic data. Various examples of data with simple to complex structures are brought in to illustrate the proposed methods.

1. Introduction

Neither the concept of symbolic data analysis (SDA) nor the method of matrix visualization (MV) are commonly understood by statisticians and users of statistical and data analysis, and we use simple examples to convey them.

1.1. Symbolic data analysis (SDA)

Symbolic data analysis (SDA: Diday, 1987; Bock and Diday, 2000; Billard and Diday, 2003) was introduced and developed for handling data with hierarchical or dependent nature and for handling large datasets. The major difference between symbolic and conventional data is that cells (rows by columns) in a data table might contain an interval, a histogram, or even

* Corresponding author. Tel.: +886 2 27835611 407; fax: +886 2 2783 1523.
E-mail address: cchen@stat.sinica.edu.tw (C.-h. Chen).

http://dx.doi.org/10.1016/j.csda.2014.04.012
0167-9473/

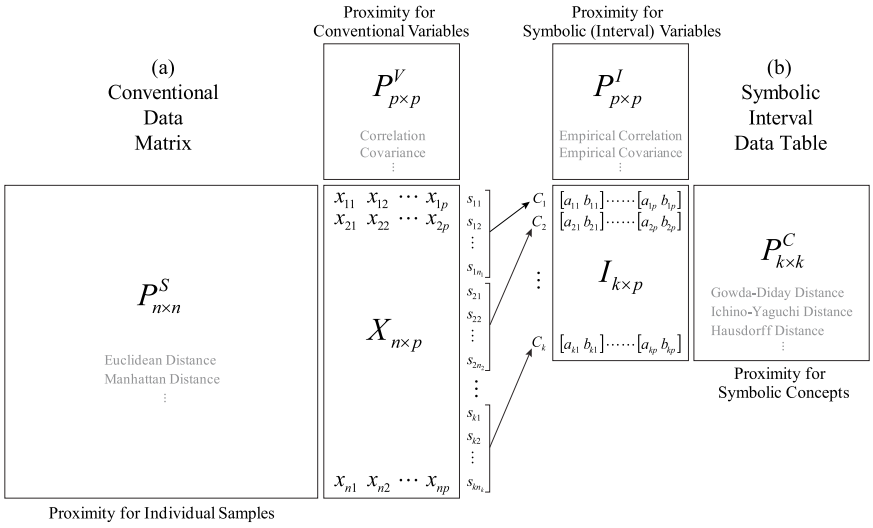

Fig. 1. Diagram for related conventional data matrix and symbolic (interval type) data table with their corresponding proximity matrices for samples/concepts and variables.

a distribution. As illustrated in Fig. 1 using interval type symbolic data, each row (sample) in the conventional data matrix $X_{n\times p}$ contains a vector of numeric values, $s_i = (x_1, \ldots, x_n)$, while each row in the symbolic data table $I_{k\times p}$, called a CONCEPT (or UNIT), contains a vector of intervals (ranges), $c_j = ([a_{j1}, b_{j1}], [a_{j2}, b_{j2}], \ldots, [a_{jp}, b_{jp}])$, that describe the behavior of a group of samples, $\{s_{j1}, s_{j2}, \ldots, s_{jp}\}$, of the variables from $X_{n\times p}$. It is also possible that instead of representing the aggregation of individual (single-valued) samples, symbolic variables describe units for which some internal variability is naturally present. In general, a symbolic data variable describes related behavior of groups of samples using ranges, histograms, or distributions for corresponding conventional variables. According to the type of symbolic data used, a statistical procedure for conventional data analysis has to be adapted to symbolic data analysis accordingly. In this we mean to include descriptive statistics, graphics and visualization, exploratory data analysis, statistical models, and inferences. Bock and Diday (2000), Billard and Diday (2003, 2006), and Diday and Noirhomme-Fraiture (2008) provide basic and necessary descriptions for these conversions from conventional data analysis to symbolic data. Among multi-valued, modal multi-valued, interval-valued, and histograms, interval data is by far the most popular symbolic data type in the literature and the most commonly encountered in practice. Many methods for interval data have been proposed, including principal component analysis (Chouakria et al., 2000; Palumbo and Lauro, 2003; Gioia and Lauro, 2006; Hamada et al., 2008), clustering analysis (Bock, 2002; Brito, 2002; Souza and de Carvalho, 2004; El Golli et al., 2004; de Carvalho et al., 2006; Chavent et al., 2006; Bock, 2008), discriminant analysis (Lauro et al., 2000; Duarte Silva and Brito, 2006), regression models (Billard and Diday, 2000; Lima Neto and De Carvalho, 2008, 2010), and multidimensional scaling (Denáux and Masson, 2000; Groenen et al., 2006; Minami and Mizuta, 2008). The crossed clustering algorithm (Verde and Lechevallier, 2005), which simultaneously groups cases and variables for symbolic data, is related to the proposed method that visualizes clustering pattern of symbolic concepts with grouping structure of interval variables in the same matrix visualization.

Two and three dimensional hypercubes came naturally for visualization of interval type symbolic data. Zoom star (Noirhomme-Fraiture and Rouard, 2000) and parallel-coordinate-plots (Lauro and Palumbo, 2003) represent each interval value (range) on a radial axis and parallel axis, respectively. Convex hulls and closed shapes were proposed for visualizing multivariate interval data (Irpino et al., 2003). Another type of visualization commonly used for clustering of symbolic data is the pyramid tree structure (Bertrand and Diday, 1985; Brito, 2002). Those aforementioned dimension reduction techniques can also be considered part of visualization processes. The most relevant symbolic data visualization work to the current study is the aggregated visual representations with the Zoomable Adjacency Matrix Explorer (ZAME) proposed in Elmqvist et al. (2008). SODAS (an academic freeware, Diday, 2002) and SYR (a commercial product, http://www.syrokko.com/) both provide various symbolic data statistical analysis algorithms and visualization tools for SDA users. Most graphical and visualization approaches for exploring symbolic data structure inherit advantages of their counterparts for conventional data, but also their disadvantages. Hypercubes (scatter-plot for symbolic data) are useful for visualizing two or three variables; histograms on parallel-coordinate-plots (box-plot for symbolic data) do not provide interactions between variables; zoom star (radar-plot for symbolic data) works for only a few concepts, and requires extensive effort to extract overall information from many concepts.

1.2. Matrix visualization (MV)

In this study we propose to use matrix visualization (MV, Chen, 2002; Chen et al., 2004; Ghoniem et al., 2005; Henry and Fekete, 2006; Wu et al., 2008; Liiv, 2010) for visualizing and clustering symbolic data using interval data as an example. We

hope to remedy some of the disadvantages in existing graphics and visualization for symbolic data. Most of the related MV techniques can be easily adapted for such other symbolic data types as multi-valued, intervals, and histograms.

MV is a graphical technique for EDA that can simultaneously explore the associations of up to thousands of subjects, variables, and their interactions, without dimension reduction. Bertin (1967) pioneered this technique by introducing reorderable matrices to systematically present data structures and relationships among samples and variables (see also de Falguerolles et al., 1997). Minnotte and West (1998) produced the software of Data Image that inherited many functions of color histogram introduced by Wegman (1990). Related techniques have been extended to outlier detection (Marchette and Solka, 2003) and time series data visualization (Roger, 2008) and have been widely used in visualization for gene expression profiles generated from microarray experiments (Eisen et al., 1998). Friendly (2002) introduced corrgrams for understanding variable association structure in correlation and covariance matrices. Chen (2002) integrated clustering and visualization of the data matrix and two proximity matrices into generalized association plots (GAP). Ghoniem et al. (2005) studied the matrix-based representations of graphs and node-link diagrams. Henry and Fekete (2006) integrated adjacency matrix visualization with network visualization. Wilkinson and Friendly (2009) had a review work on the history of the cluster heat map, while Liiv (2010) gave a summary of seriation and matrix reordering methods. The cluster heat map (matrix visualization) has been called a post-genomic visual icon in Weinstein (2008).

We use the GAP (Chen, 2002) approach to illustrate the basic principles of MV for conventional data. As illustrated in Fig. 1(a), given a data matrix $X_{n\times p}$ with n rows (samples) and p columns (variables), proximity matrices are computed for measuring relationships among samples ($P^S_{n\times n}$) and association structures between variables ($P^V_{p\times p}$). Various seriation (reordering) algorithms (Tien et al., 2008; Wu et al., 2010; Liiv et al., 2012) can then be applied to sort the proximity matrices so that groups of variables and clusters of samples can be formed along the main diagonals of $P^V_{p\times p}$ and $P^S_{n\times n}$. The data matrix $X_{n\times p}$ is then two-way sorted using the orders of variables and samples from reordered $P^V_{p\times p}$ and $P^S_{n\times n}$ so that interaction patterns of sample clusters on variable groups embedded in $X_{n\times p}$ can be identified. The reordered $X_{n\times p}$ and proximity matrices $P^V_{p\times p}$ and $P^S_{n\times n}$ are then displayed as matrix maps through suitable color projections so that sample-clusters, variable-groups, and interactions between the two sides can be visually extracted. Various extensions of GAP for analyzing binary data, categorical data, data with cartography structure, and other MV related techniques have also been developed. A Java version of GAP developed by Wu et al. (2010) is available for public use: http://gap.stat.sinica.edu.tw/Software/index.htm.

We describe the basic principles of MV for symbolic data using a common toy dataset in Section 2. Section 3 illustrates proposed MV techniques for interval symbolic data in two examples. Several more advanced MV methods uniquely developed for symbolic data are discussed in Section 4 using the same examples. Discussion and concluding remarks are given in Section 5.

2. Matrix visualization for interval (range) type symbolic data

It is necessary for us to convert MV techniques, of clustering and visualizing conventional continuous data, to interval symbolic data. The conversion involves the computation of a proximity matrix for interval variables, $P^I_{p\times p}$, the computation of a proximity matrix for symbolic concepts, $P^C_{k\times k}$, and the color coding of an interval data table, $I_{k\times p}$ (Fig. 1(b)). We propose our methods for doing these tasks in the next three subsections. We borrowed suitable proximity measurements from the literature, and have developed new color coding schemes for representing interval type data. Color coding of proximity matrices here is similar to that used for conventional data.

2.1. Proximity matrix for interval (range) variables

There are not many choices for measuring association between interval type symbolic data variables. We adopt the empirical covariance function and empirical correlation coefficient for interval data proposed in Billard and Diday (2006), and Duarte Silva and Brito (2006). Let $I_i = (I_{1i}, I_{2i}, \ldots, I_{ki})^T$ be the interval data for the ith variable with k concepts, where $I_{ci} = [a_{ci}, b_{ci}]$, $c = 1, 2, \ldots, k$. The *empirical covariance function* between I_i and I_j is

$$Cov(I_i, I_j) = \frac{1}{4k}\sum_{c=1}^{k}[(a_{ci}+b_{ci})(a_{cj}+b_{cj})] - \frac{1}{4k^2}\left[\sum_{c=1}^{k}(a_{ci}+b_{ci})\right]\left[\sum_{c=1}^{k}(a_{cj}+b_{cj})\right]. \tag{1}$$

The *empirical correlation coefficient* between I_i and I_j is

$$r(I_i, I_j) = \frac{Cov(I_i, I_j)}{S_{Z_i}S_{Z_j}}, \tag{2}$$

where

$$S^2_{Z_i} = \frac{1}{3k}\sum_{c=1}^{k}(b^2_{ci} + b_{ci}a_{ci} + a^2_{ci}) - \frac{1}{4k^2}\left[\sum_{c=1}^{k}(b_{ci}+a_{ci})\right]^2.$$

Table 1
Distance measures for interval type symbolic data proposed in Billard and Diday (2006).

Measure name	Formula	Component detail
The Gowda–Diday distance (Gowda and Diday, 1991)	$\sum_{r=1}^{p} D(I_{ir}, I_{jr})$	$D(I_{ir}, I_{jr}) = \frac{\lvert a_{ir}-a_{jr}\rvert}{\lvert \max_c b_{cr}-\min_c a_{cr}\rvert} + \frac{\lvert\lvert b_{ir}-a_{ir}\rvert-\lvert b_{jr}-a_{jr}\rvert\rvert}{\max(b_{ir},b_{jr})-\min(a_{ir},a_{jr})} + \frac{\lvert b_{ir}-a_{ir}\rvert+\lvert b_{jr}-a_{jr}\rvert-2I_r}{\max(b_{ir},b_{jr})-\min(a_{ir},a_{jr})}$ where $I_r = \lvert \max(a_{ir}, a_{jr}) - \min(b_{ir}, b_{jr})\rvert$
The Ichino–Yaguchi distance (Ichino, 1988)	$\sqrt[q]{\sum_{r=1}^{p} D(I_{ir}, I_{jr})^q}$	$D(I_{ir}, I_{jr}) = \lvert[a_{ir}, b_{ir}] \cup [a_{jr}, b_{jr}]\rvert - \lvert[a_{ir}, b_{ir}] \cap [a_{jr}, b_{jr}]\rvert + \gamma(2\lvert[a_{ir}, b_{ir}] \cap [a_{jr}, b_{jr}]\rvert - \lvert[a_{ir}, b_{ir}]\rvert - \lvert[a_{jr}, b_{jr}]\rvert)$ where $0 \le \gamma \le 0.5$
The L_1 distance	$\sum_{r=1}^{p} D(I_{ir}, I_{jr})$	$D(I_{ir}, I_{jr}) = \lvert \frac{a_{ir}+b_{ir}}{2} - \frac{a_{jr}+b_{jr}}{2}\rvert$
The L_2 distance (de Carvalho et al., 2006)	$\sum_{r=1}^{p} D(I_{ir}, I_{jr})$	$D(I_{ir}, I_{jr}) = (\frac{a_{ir}+b_{ir}}{2} - \frac{a_{jr}+b_{jr}}{2})^2$
The City-Block distance (Souza and de Carvalho, 2004)	$\sum_{r=1}^{p} D(I_{ir}, I_{jr})$	$D(I_{ir}, I_{jr}) = \lvert a_{ir} - a_{jr}\rvert + \lvert b_{ir} - b_{jr}\rvert$
The Hausdorff distance (Chavent and Lechevallier, 2002)	$\sum_{r=1}^{p} D(I_{ir}, I_{jr})$	$D(I_{ir}, I_{jr}) = \max(\lvert a_{ir} - a_{jr}\rvert, \lvert b_{ir} - b_{jr}\rvert)$
The Euclidean Hausdorff distance	$\sqrt[2]{\sum_{r=1}^{p} D(I_{ir}, I_{jr})^2}$	$D(I_{ir}, I_{jr}) = \max(\lvert a_{ir} - a_{jr}\rvert, \lvert b_{ir} - b_{jr}\rvert)$
The normalized Euclidean Hausdorff distance	$\sqrt[2]{\sum_{r=1}^{p} [\frac{D(I_{ir}, I_{jr})}{H_r}]^2}$	$D(I_{ir}, I_{jr}) = \max(\lvert a_{ir} - a_{jr}\rvert, \lvert b_{ir} - b_{jr}\rvert)$ $H_r^2 = \frac{1}{2k^2} \sum_{i=1}^{k} \sum_{j=1}^{k} D(I_{ir}, I_{jr})^2$
The span normalized Euclidean Hausdorff distance	$\sqrt[2]{\sum_{r=1}^{p} [\frac{D(I_{ir}, I_{jr})}{\lvert R_r\rvert}]^2}$	$D(I_{ir}, I_{jr}) = \max(\lvert a_{ir} - a_{jr}\rvert, \lvert b_{ir} - b_{jr}\rvert)$ $\lvert R_r\rvert = \max_c b_{cr} - \min_c a_{cr}$

Both the empirical covariance and correlation measures rely on an assumption of uniformity within each observed interval. The empirical correlation coefficient for interval data is bounded by −1 and 1, and reduces to the Pearson product-moment correlation when intervals are trivial (single points). Refer to Duarte Silva and Brito (2006) for more properties of the empirical correlation coefficient and the empirical covariance function.

2.2. Distance matrix for concepts with interval variables

Several proposals have been made for measuring the among-concept relationship for interval type SDA variables. For two concepts $C_i = [I_{i1}, I_{i2}, \ldots, I_{ip}]$ and $C_j = [I_{j1}, I_{j2}, \ldots, I_{jp}]$ with p interval variables, we summarize several distance measures proposed by Billard and Diday (2006) in Table 1.

Some distance measurements for interval variables and concepts do not have desirable properties, and it is expected that others will be proposed. The L1 and L2 distances utilize only the midpoints of the intervals; the choice of parameter γ may not be suitable for all $C(k, 2)$ pairs of between concept Ichino–Yaguchi distance; the City-Block distance sums up marginal differences of the two end points; the four Hausdorff distances combine information of the two end points with different scaling factors; the Gowda–Diday distance takes all aspects of interval data into consideration and for sure will identify very similar concepts (concepts that are close to each other on all aspects) but may have problems on grouping concepts that do not share all common properties. For now, we accept available distances for representing between-variable and among-concept relationships for interval data, and proceed to the matrix visualization process. Throughout, the empirical correlation coefficient (2) and the span normalized Euclidean Hausdorff distance are employed for computing the between-variable and among-concept distance matrices $P^V_{p\times p}$ and $P^C_{k\times k}$, respectively. Of course choices on the between-variable and between-concept measures result in different distance matrices and therefore permutations of distance matrices with data matrix. The GAP software for classical data places Euclidean distance as the default between-sample measurement for its intuitive interpretation, easy computation, and good practice. We also choose among the set of Euclidean Hausdorff distances as the default between-concept distance for interval data for similar reasons. Because our approach of matrix visualization computes all $C(k, 2)$ pairs of between-concept distance using all p interval variables the span normalized Euclidean Hausdorff distance stands out for its robustness when dealing with interval variables of varying scales. Choice of between-concept distance does not have significant impact for standardized variables or variables with common scales. Matrix visualization allows users to visualize the permuted data matrix with two distance matrices so users are encouraged to practice with various measurements and permutation algorithms to have a better understanding of data structure for analysis.

We use the Bats data (Billard et al., 2009), see Fig. 2, to compare the following between-concept distances: the Gowda–Diday distance (GD), the L1 distance (L1), the L2 distance (L2), the City-Block distance (CB), the Hausdorff distance (HD), the Euclidean Hausdorff distance (EHD), the normalized Euclidean Hausdorff distance (nEHD), and the span normalized Euclidean Hausdorff distance (snEHD). There are 21 species of bats described by 4 interval variables (Head, Tail, Height and Forearm range). Fig. 2(a) has MVs of these 8 distance matrices each individually sorted by the HCT–R2E algorithm to be introduced in Section 2.4. Roughly we have the following observations: the GD matrix is the most distinct one among the 8 matrices; L1, CB, and L2 matrices have similar patterns, in particular L1 and CB are very similar to each other; HD,

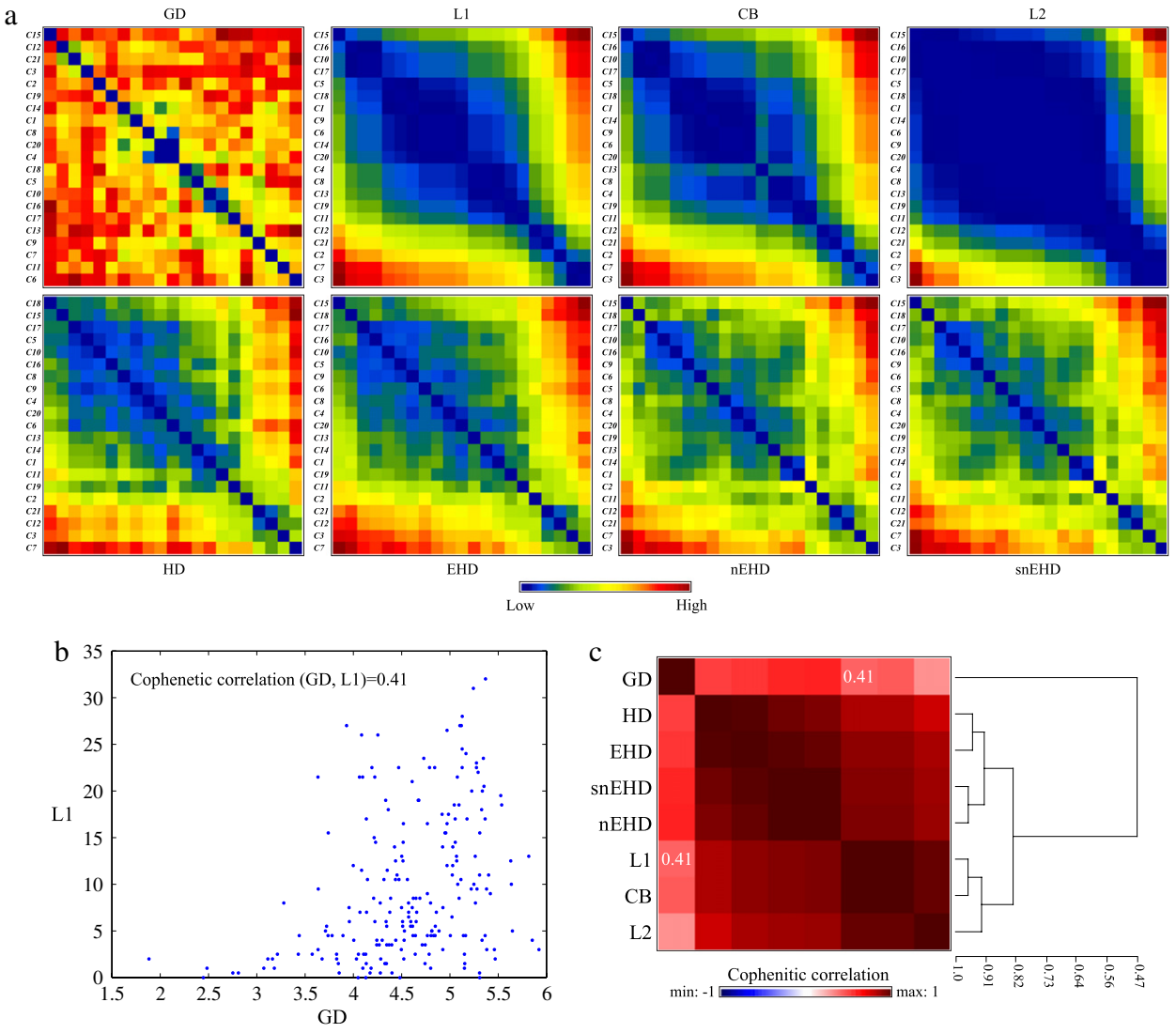

Fig. 2. Comparison of between-concept distance measures. (a) Matrix visualization of 8 distance matrices each individually sorted by the HCT–R2E algorithm. (b) Cophenetic correlation of the Gowda–Diday distance (GD) matrix and the L1 distance matrix. (c) Matrix visualization and clustering of the pairwise cophenetic correlation among the 8 distance matrices. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

EHD, nEHD, and snEHD matrices form a Hausdorff distance group where nEHD and snEHD are almost identical. In order to have a more rigorous comparison we adopt the concept of cophenetic correlation coefficient (Sokal and Rohlf, 1962). Cophenetic correlation of two matrices treats all lower-triangle elements of the two matrices as two vectors and computes their Pearson correlation. Fig. 2(b) displays the corresponding $C(21, 2) = 210$ lower-triangle elements of GD and L1 distance matrices with their cophenetic correlation computed as 0.41. Fig. 2(c) has the MV for all $C(8, 2) = 28$ pairwise cophenetic correlations among the 8 distance matrices sorted by the dendrogram of HCT–R2E. Structures from both the MV and the dendrogram reconfirm the previous visual observations of Fig. 2(a).

2.3. Color coding for interval (range) data table

It remains to color code a data table with p interval variables and k concepts. Elmqvist et al. (2008) proposed eight different glyphs for representing aggregated data structures. Among them Min/max (histogram), Min/max (band) and Min/max (tribox) are related to interval data. Instead of using color bands to represent intervals as in this study it is possible to use these three different glyphs to display interval data. Saito et al. (2005) introduced a two-tone pseudo coloring technique for large-scale one-dimensional data that is useful for displaying symbolic data where original un-aggregated data or distribution of data is known. The Bats data in Fig. 2 is used again to illustrate our proposed coloring technique. We first

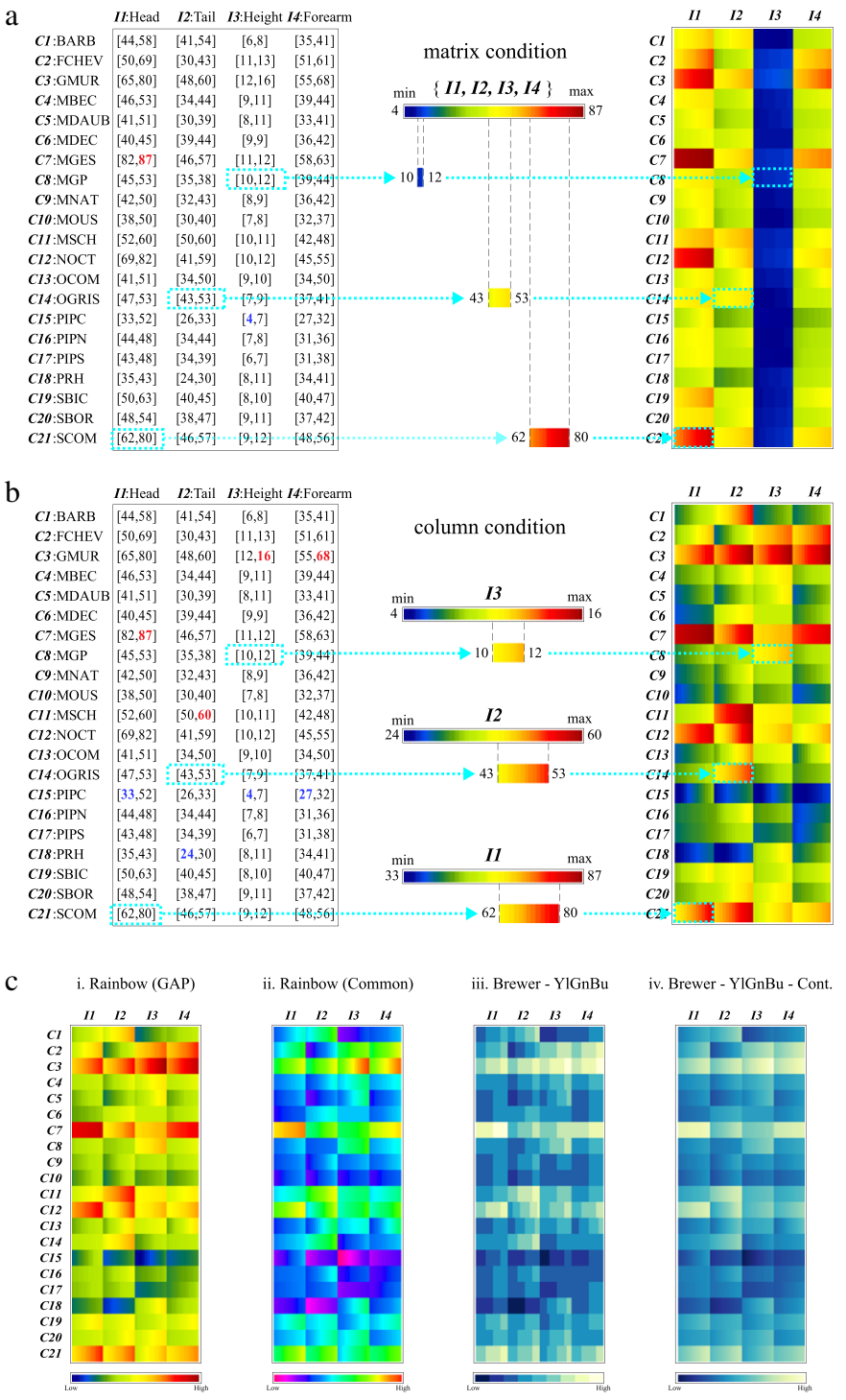

Fig. 3. Color-coding scheme for interval-valued symbolic data using the Bats example. (a) Matrix condition. (b) Column condition. (c) Standardized data with different color maps.

identify the minimum and maximum values of the 21×4 data table as 4, the lower bound of $I_{15:3}$, and 87, the upper bound of $I_{7:1}$. A rainbow color spectrum is employed to represent the range (4, 87). This rainbow color map used in the GAP system is different from the commonly used rainbow color map that is considered harmful in several aspects (Borland and Taylor, 2007). Each of the 84 intervals is then represented by its corresponding segment within the complete spectrum to form the integrated interval data table with 21 concepts on 4 interval variables seen in Fig. 3(a), $I_{21\times4}$.

From the figure, one can see that variable I_3 for Height has relatively smaller sizes and shorter ranges than the other three variables, since the whole column is coded with dark blue segments. This is akin to seeing an outlier or an outlying group in a scatter-plot in which the within-group resolution is sacrificed to show the between group structure. We suspect the variable "Height" should be height of ear and the unit for this variable is cm, while it is mm for the other three. We do not know if this is correct but have developed the column-condition color coding for resolving this color space resolution problem in Fig. 3(b). Instead of using a color spectrum to represent the whole matrix range of (4, 87) as in Fig. 3(a) for the matrix-condition, the column-condition applies the same color spectrum to display four column ranges of (33, 87), (24, 60), (4, 16), and (27, 68) for the four interval variables respectively in Fig. 3(b). We see each column now has better color resolution in describing the within-variable structure while the between-variable structure no longer exists. Instead of using column condition to cope with different ranges observed among interval variables, it is possible to standardize (Guo et al., 2012) the interval symbolic data first before color-coding them with the matrix condition. The effect of standardization in Fig. 3(c-i) is very similar to that of column condition in Fig. 3(b).

Selection of color spectrum (map) may also introduce different visual perceptions. We shall illustrate the color map effect in Fig. 3(c-ii–iii–iv) using three more different color maps. The first one is the commonly used rainbow color map that is considered harmful. One of the critiques is about the closeness of the red and purple ends in the color space makes it difficult to visually distinguish values from the two extremes as can be seen in Fig. 3(c-ii). The second one is the YlGnBu (yellow–green–blue) color map from Brewer's color map system (Brewer, 1994, 1999) recommended by most of the visualization literature (Rosenberg, 2003; Ware, 2004; Elmqvist et al., 2011; Micallef et al., 2012). The original Brewer's color map with only 3–9 segments may not be suitable for displaying interval data because the segmented pattern can be confounded with the interval data structure as in Fig. 3(c-iii). The continuous version (Wijffelaars et al., 2008) of Brewer's color maps will do better jobs for interval data Fig. 3(c-iv).

The proposed GAP rainbow color map improves over the standard rainbow color map but has not been validated experimentally. There is no such a universally better color map for all types of data structure. It is important for the users of color map to understand well their data structure and properties before choosing the appropriate one for presenting their data. It is also necessary for the viewers to consult the color map legend when there is the need. In *i*GAP users have the option to try different color maps including the GAP rainbow map and those from the Brewers color map system.

2.4. Seriation/clustering and matrix visualization

We now proceed with MV as with conventional data and perform the seriation (reordering) steps. We compute empirical correlations for $P^I_{p\times p}$ and span normalized Euclidean Hausdorff distances for $P^C_{k\times k}$. The GAP approach matrix visualization is an environment for data visualization, not a direct clustering method. When suitable permutations for rows (concepts) and columns (variables) are identified natural clustering of concepts and grouping of variables will emerge in the data matrix and two proximity matrices. When hierarchical tree seriation is used the branching structure of the dendrogram can be used to cluster concepts and variables. Throughout this article the rank-two ellipse-guided hierarchical clustering tree seriation (HCT–R2E, Tien et al., 2008) is used to permute the rows (concepts) and columns (variables). The span normalized Euclidean Hausdorff distance matrix of the Bats data (Billard and Diday, 2006) is employed to illustrate the HCT–R2E seriation algorithm. The HCT–R2E method has the base of an average linkage hierarchical clustering tree (HCT) to guarantee a coherent local pattern with blue to dark green dots (smaller distances) concentrating around the main diagonal, Fig. 4(a). A regular HCT does not necessarily maintain good global structure so the orange to dark red dots (larger distances), between the two main clusters, still stay near the main diagonal in Fig. 4(a). The rank-two ellipse seriation (R2E, Chen, 2002) modified from singular value decomposition on the other end always provides a smooth global structure but not necessarily the coherent local pattern, Fig. 4(b). The dendrogram in Fig. 4(a) has 21 concepts with 20 intermediate nodes and each of them can be flipped independently resulting in 2^{20} possible orderings of the terminal nodes from the same HCT and identical distance matrix. The R2E ordering (Fig. 4(b)) is then used to guide the 20 intermediate nodes (Fig. 4(c)) to flip (red nodes) or not to flip (black nodes). The resulting permutation thus simultaneously identifies coherent local patterns with smooth global structures in Fig. 4(c).

3. Examples

Many matrix visualization techniques and tools have been developed for analyzing interval type symbolic data. We use two examples with interval data for demonstrating some of them.

3.1. Meteorological stations in China

The Long-Term Instrumental Climatic Database of the Peoples Republic of China has been widely used in the SDA literature when proposing statistical procedures of interval type symbolic data, such as standardization of variables (Guo et al., 2012), clustering analysis and self-organizing maps (Verde et al., 2003; El Golli et al., 2004; Chavent et al., 2006), and MLE and MANOVA (Brito and Duarte Silva, 2012). Various versions of this dataset have been employed. Here we consider the lowest and highest temperature observed over the twelve months of 1988 at sixty meteorological stations in China. The

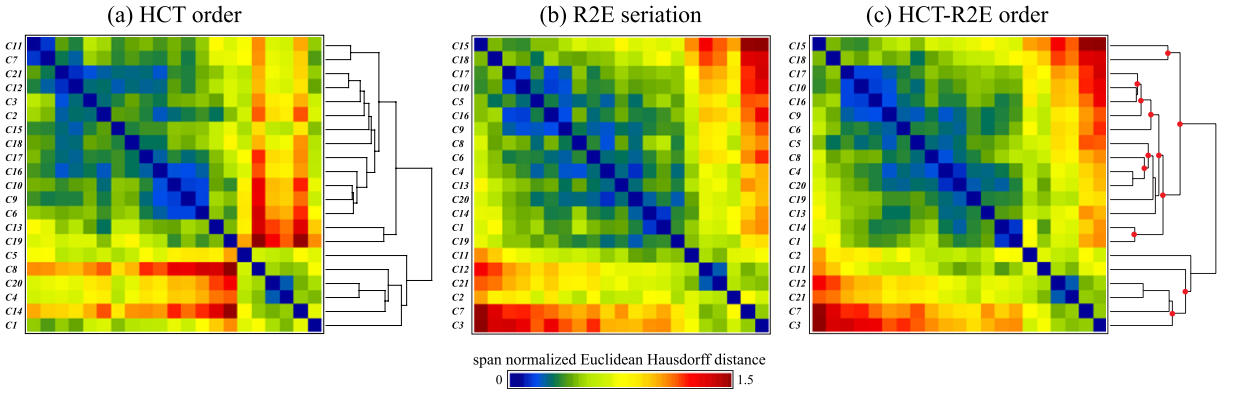

Fig. 4. The HCT–R2E algorithm for the span normalized Euclidean Hausdorff distance matrix of the Bats data (Billard and Diday, 2006). (a) Matrix visualization with hierarchical clustering tree (HCT). (b) Matrix visualization with rank-two ellipse seriation (R2E). (c) Matrix visualization with R2E guided HCT (HCT–R2E); red dots on the dendrogram indicate intermediate nodes with flips. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

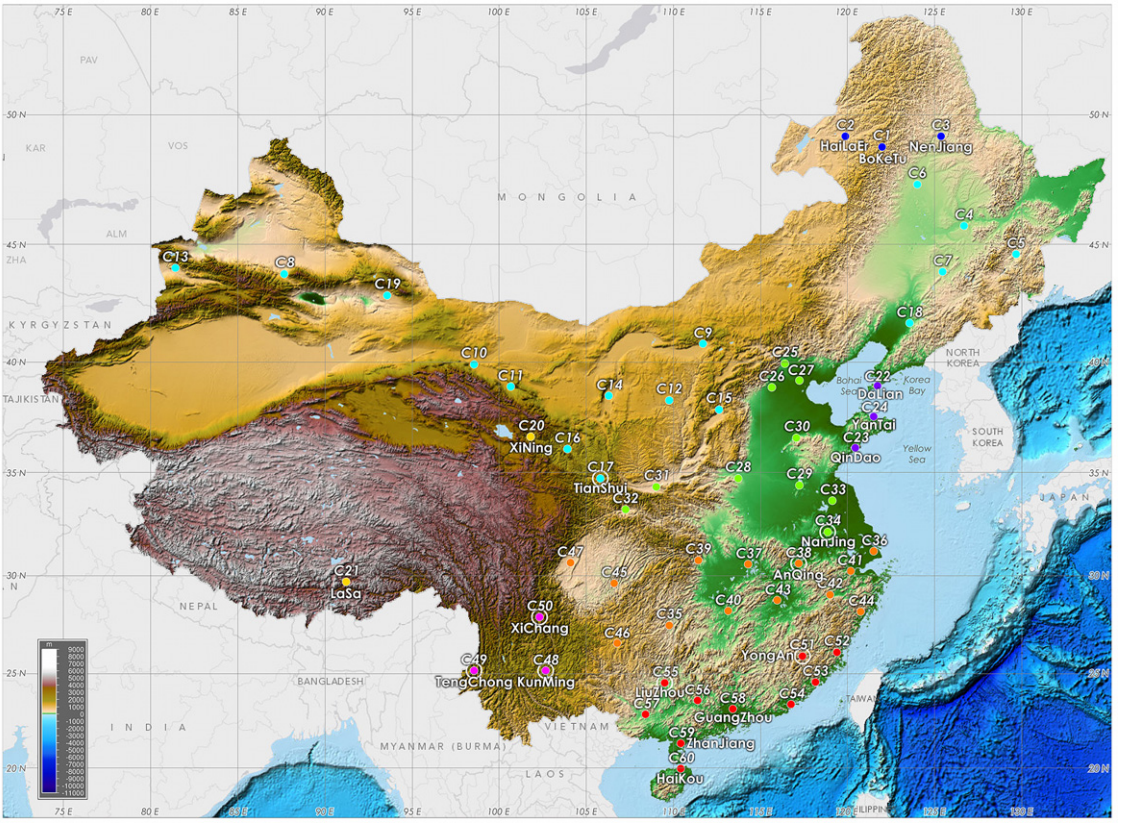

Fig. 5. Sixty China meteorological stations in an elevation map. Colors for representing related clusters of stations identified from dendrogram structure in Fig. 6(a) are used to code each of the individual stations and white outer circle for those stations with number of disagreements $\geq$48. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)
Source: NOAA web page.

data table has 720 intervals (ranges) for 60 stations (concepts) on 12 months (January–December). These data are from the Research Data Archive (RDA) which is maintained by the Computational and Information Systems Laboratory (CISL) at the National Center for Atmospheric Research (NCAR). NCAR is sponsored by the National Science Foundation (NSF). The original data are available from the RDA (http://dss.ucar.edu) in dataset number ds578.5.

Fig. 5 shows the locations of these sixty stations on a China elevation map modified from an original map obtained from the NOAA (National Oceanic and Atmospheric Administration of the United State Department of Commerce) web page: http://www.noaa.gov/. Fig. 6(a) displays the three sorted matrix maps: the 60 stations by 12 months temperature

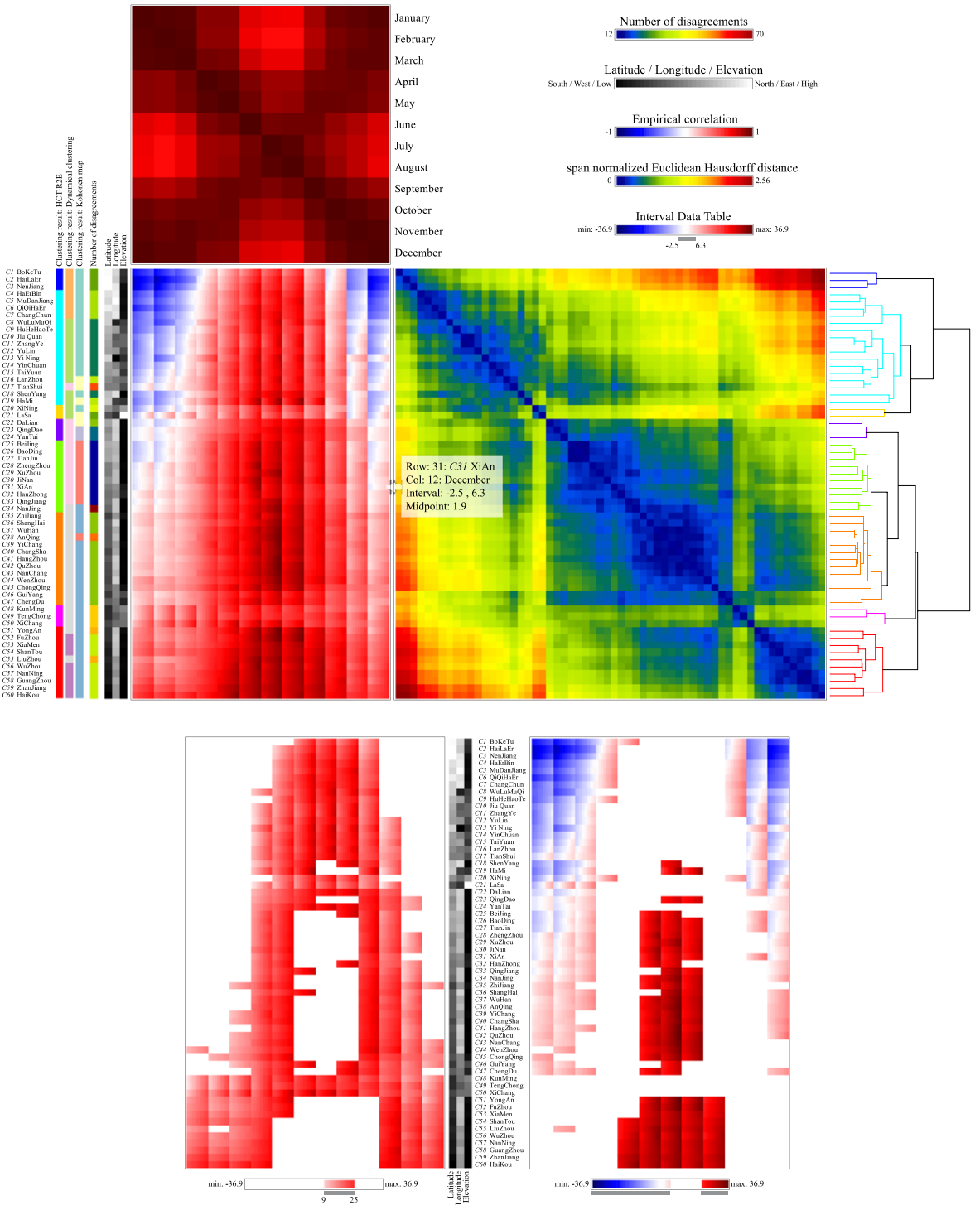

Fig. 6. (a) Three MV maps sorted by HCT–R2E dendrograms for 12 monthly temperature range variables on 60 China meteorological stations data. (b) Midpoint condition range display for 60 China meteorological stations data in (a). Left panel: only temperature intervals with midpoint within the range of (9–25 °C) are displayed; right panel: only intervals with midpoint outside the range of (9–25 °C) are displayed.

interval data table, $I_{60\times 12}$, the empirical correlation map for 12 months, $P^{I}_{12\times 12}$, and the span normalized Euclidean Hausdorff distance map for 60 stations, $P^{C}_{60\times 60}$. While we sorted the 60 stations using HCT–R2E seriation, the natural order of 12 months allows for easier interpretation. A symmetric blue (negative in Celsius degree) to white (zero) to red (positive) color spectrum is used to represent the range (−36.9–36.9 °C), while the range of temperature observed in the data table is

(−29.6–36.9 °C). Next to $I_{60\times 12}$ are three covariates, represented by a gray (white–black) spectrum for longitude (west–east), latitude (north–south), and elevation (high–low) of the 60 stations. Clustering results of the HCT–R2E method, the dynamical clustering method (Diday, 1971; Chavent and Lechevallier, 2002), and the Kohonen map method (Bock, 2008) are displayed as three discrete covariates with distinct color sets for easier comparison. Corrected Rand Index (Hubert and Arabie, 1985), ranges from 0 (distinct) to 1 (identical), is calculated to measure similarity between clustering results among the three methods as (HCT–R2E vs. Dynamical, HCT–R2E vs. SOM, Dynamical vs. SOM) = (0.55906, 0.4617, 0.5292). So HCT–R2E and Dynamical clustering produce more similar clustering structure while HCT–R2E and SOM give most distinct grouping patterns. We further decompose this Corrected Rand Index into number of disagreements contributed by each sample (station). Summation of three numbers of disagreements from three paired comparisons for each station is displayed as another continuous covariate in Fig. 6(a). The following eight stations with largest total number of disagreements among the three clustering methods are indexed by a white circle in Fig. 5 (C_{34}:NanJing, C_{17}:Tain Shui, C_{38}:AnQing, C_{51}:YongAn, C_{55}:LiuZhou, C_{49}:TengChong, C_{50}:XiChang, and C_{48}:KunMing). While C_{48}:KunMing, C_{49}:TengChong, and C_{50}:XiChang have relatively higher elevations the other five stations are all located in the geographical boundaries of various clusters.

While these numerical indices do provide readers some ideas about the similarity and difference between clustering methods and among different samples, readers should consult the matrix visualization of data profiles in Fig. 6(a) for a more comprehensive understanding about the mechanism behind these clustering algorithms. Users can also click on these clustering covariates to sort the data table and distance matrix accordingly for better understanding and interpretation of related clustering mechanisms. In $P^{I}_{12\times 12}$ one sees that the interval variables for 12 months are positively correlated (in the Empirical Correlation sense) with consecutive months showing stronger correlations, especially in the winter.

Instead of many clusters of stations each with different temperature patterns, as reported in other articles, the most prominent visual pattern from both sorted $I_{60\times 12}$ and $P^{C}_{60\times 60}$ is a rather smooth trend (patterns of temperatures and the span normalized Euclidean Hausdorff distances in color representations) from the coldest stations in the north (C_1:BoKeTu, C_2:HaiLaEr, C_3:NenJiang) to the warmest stations in the south (C_{60}:HaiKou, C_{59}:ZhanJiang, C_{58}:GuangZhou). The color coded four main branches of the dendrogram (cyan, green, orange, red) illustrate four major groups of stations from north (cold) to south (warm), as reported elsewhere, with four minor groups of stations that do not merge well into either the main smooth trend in color or the main dendrogram architecture. It is possible to split major clusters into smaller and more coherent sub-clusters as desired.

Here we discuss further those four minor groups of stations from north to south. The group of northern-most stations (BoKeTu, HaiLaEr, NenJiang) in blue differ (in temperature pattern, distance structure, and dendrogram branching) from the rest of the 57 stations due to their colder and longer winter weather. The stations C_{20}:XiNing and C_{21}:LaSa in yellow came next, with rather different temperature patterns (in $I_{60\times 12}$) and distance structure (in $P^{C}_{60\times 60}$) from neighboring stations, potentially because of their relatively high elevations, as can be seen from the lighter color in the covariate spectrum and from Fig. 5. Below XiNing and LaSa is the group (C_{22}:DaLian, C_{23}:QinDao, C_{24}:YanTai) in purple that branches out of the two major groups of cyan and green in the dendrogram. These northern stations do not suffer from cold winter temperatures because weather is moderated by the Bohai Gulf (bounded by two peninsulas) and the Korea Bay within the Yellow Sea. DaLian is China's northernmost major ice-free seaport also due to its location near the Bohai Gulf. The bottom group of outliers (C_{48}:KunMing, C_{49}:TengChong, C_{50}:XiChang) in magenta have consistent medium range temperature intervals across the year, which may also be due to their relatively higher elevations.

With the assistance of MV in Fig. 6 for the sorted 60 stations in the meteorological station data table, users can better understand why certain stations are classified together or not because they share similar or distinct 12-month temperature profiles. Matrix visualization can visually extract overall patterns for up to thousands of units and variables simultaneously, but one may not be able to identify the precise value or range for individual cells (intervals in the present study). The *i*-function (information retrieve) in developed software is provided for remedying this shortcoming of MV. When users click on any cell (interval), for example C_{31}:XiAn/December in Fig. 6(a), detailed information for that cell (interval) is displayed in a pop-out window with exact segment (−2.5–6.3) of color spectrum signified in the color legend.

3.2. Japan Minryoku 2010 data

In this subsection we apply the proposed MV techniques on the Minryoku (Manpower in Japanese) 2010 data published by Asahi Shimbun Publications Inc. (Minryoku 2010, 2010 DVD with web page https://minryoku.jp/enduser/). This database provides census-like manpower information and economic activities for four levels of hierarchy of townships (Level 1: 10 regions; Level 2: 151 areas; Level 3: 821 districts; Level 4: 1899 cities) in Japan from 1989 to 2010. For our purposes we treat each Level 2 area as one concept, with each concept to include 2–56 (median 10) Level 4 cities (individual sample). Due to potential outlier effects and scaling problems we used ranks, rank 1 represents the smallest numeric value and rank 1899 stands for the largest. It is possible to standardize variables with varying scales first as described in Guo et al. (2012) but the outlier effects remain. The rank-transformation may introduce the problem of over-generalization as discussed in Bock and Diday (2000, pages 95) and following, and in Diday and Noirhomme-Fraiture (2008, pages 47) and following. Here the rank-transformation is not just used to tackle both outlier effects and scaling problems but also to have a simpler demonstration of the proposed MV method. After excluding several redundant variables and variables with too many missing values from the original data with 70 variables, we further merged the original 14 population by age group (5-year interval) variables

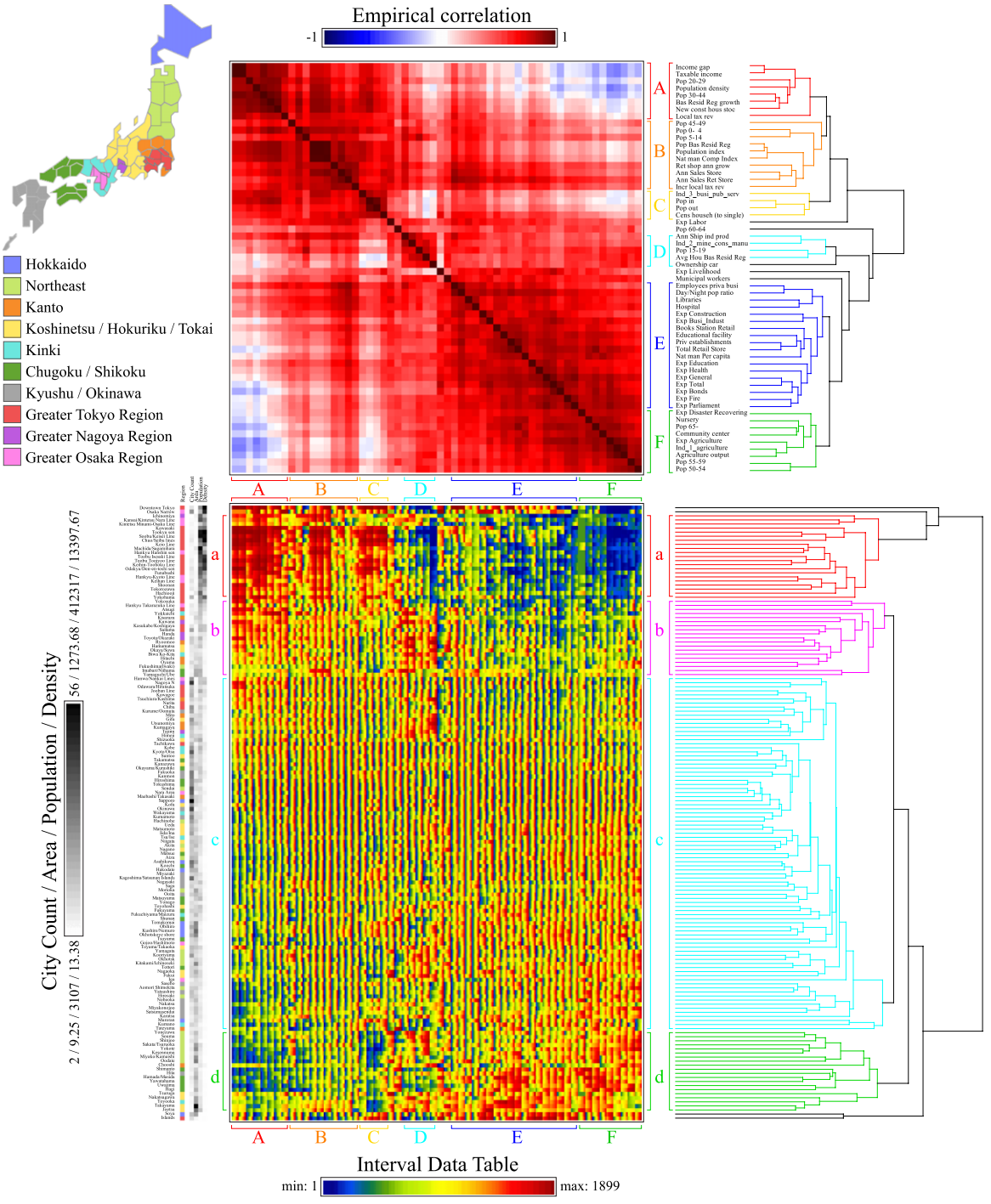

Fig. 7. Two matrix maps for Minryoku 2010 data sorted also by HCT–R2E seriations: the 151 areas by 58 rank interval table, $I_{151\times58}$, the empirical correlation map for 58 manpower rank interval variables, $P^I_{58\times58}$. The span normalized Euclidean Hausdorff distance map for 151 areas, $P^C_{151\times151}$, is not shown due to limitation of space.

into 10 variables (consecutive age groups with correlation higher than 0.9 were merged). The data table for analysis had 58 interval variables (range 1–1899) on 151 concepts (Level 2 areas) for 2010.

Fig. 7 displays two matrix maps for the Minryoku 2010 data sorted by HCT–R2E seriations: the 151 areas by 58 rank interval table, $I_{151\times58}$, the empirical correlation map for 58 manpower rank interval variables, $P^I_{58\times58}$. The span normalized Euclidean Hausdorff distance map for 151 areas, $P^C_{151\times151}$, is not shown due to limitations of space. Color coded strips for one nominal (Level 1 region) and four continuous (city counts, city area size, population size, and population density) covariates

Table 2
Information structure of the six major variable groups (with number of variable for each group) on four main area clusters (with number of areas within each cluster). Listed in each cell (area cluster by variable group) is the mean interval (minimum, maximum) for all intervals within that corresponding intersection of area cluster and variable group.

Area cluster	Variable group					
	A (8)	B (10)	C (4)	D (4)	E (18)	F (9)
a (21)	[1394, 1747]	[886, 1626]	[1178, 1638]	[368, 1089]	[244, 892]	[176, 494]
b (19)	[884, 1622]	[517, 1629]	[417, 1362]	[806, 1619]	[320, 1231]	[376, 1196]
c (87)	[254, 1468]	[182, 1690]	[253, 1573]	[318, 1617]	[278, 1686]	[440, 1664]
d (20)	[302, 837]	[349, 1315]	[272, 911]	[690, 1336]	[742, 1557]	[947, 1582]

are displayed next to the sorted data table $I_{151\times 58}$ for easier retrieval of area information. We only elaborate on some of the more significant observations one might conclude from Fig. 7. Readers can search for interesting information by examining related figures more carefully, or avail themselves of the data and software.

Here we examine Fig. 7 for the 58 variables. Using the dendrogram branching structure with color pattern of the sorted correlation matrix, $P^{l}_{58\times 58}$, we highlighted six more coherent variable groups in color and labeled them A–F as follows: (A) Income, Tax, with Main working population; (B) Population Indices; (C) Business and Public services with Migration and Census; (D) Industry and Car; (E) Stores, Education, and Expenditure; (F) Agriculture and Senior Citizen. There are five stand alone variables in addition to the six variable groups. We also summarized the clustering pattern of 151 concepts (Level 2 areas) from the dendrogram branching structure, with four main clusters of areas (a–d) having the following descriptions: (a) A cluster of urban areas from the Greater Tokyo and Greater Osaka Regions with highest population size and density; (b) The industrial areas (Toyota for Toyota Motors; Okazaki for Mitsubishi Motors; Hitachi for the Hitachi company) from seven regions; (c) The main cluster of areas with large city counts and a transitional structure from urban and industrial areas to rural areas; (d) A major cluster of rural areas, Northeast, Chugoku/Shikoku, and Koshinetsu/Hokuriku/Tokai, with high area size and low population density. There are also four outlying areas: Downtown Tokyo, Downtown Osaka, the area "Island", from the Greater Tokyo region with nine island cities, and the area "Soya", at the north tip of Hokkaido, form a cluster of two outliers.

A rainbow color spectrum is employed to code the data table with the blue for smaller and red for larger ranks. A matrix color coding is applied since all interval variables have the range 1–1899. One basic principle for visualizing structure in the sorted data table $I_{151\times 58}$ is the following: a coherent blue (red) band represents an interval with small (large) rank in a compact range, while a band with more vivid color (blue–yellow–red) stands for an interval with a larger range of ranks.

The most prominent global view in Fig. 7 is the strong opposing structures of columns (variables) and of rows (areas). The upper-left (ABC by ab) and lower-right (EF by d) corners are full of intervals with larger ranks while the upper-right (F by a) and lower-left (ABC by d) corners are occupied by intervals with smaller ranks. From the nominal covariate color spectrum of Level 1 for main regions, a smooth trend from the top rows of urban areas (Greater Tokyo Region in red, Greater Osaka Region in magenta, Greater Nagoya Region in purple) to the bottom rows occupied by rural areas (Koshinetsu/Hokuriku/Tokai in yellow, Northeast in green, Chugoku/Shikoku in light green, Kyushu/Okinawa in gray) can be identified. Four continuous covariates also have strong correlations with this trend from top to bottom.

Table 2 summarizes the information structure of the six major variable groups on four main area clusters to be seen in the sorted Minryoku data table $I_{151\times 58}$ in Fig. 7. By examining the intersections of the four area clusters (a–d) and the six variable groups (A–F) we see why certain groups of variables are positively or negatively correlated with each other (variable groups A with F, for example), why certain areas are similar or different, and how each cluster of outliers behaves differently from other clusters. Variables in group D have the most complicated distribution structure (relative to the main pattern) among the clusters of areas.

Methods for interval symbolic data, such as principal component analysis, clustering analysis, discriminant analysis, regression models, and multidimensional scaling, can be applied to the Minryoku data to extract information in the data table or for answering user questions. The sorted MV displays with dendrograms in Fig. 7 serve as an exploratory data analysis from which users might formulate more precise hypotheses and apply appropriate statistical methods/procedures. An MV display such as Fig. 7 has the potential to serve as a diagnostic in examining the results of these procedures.

4. Unique features for MV of interval symbolic data with the computation environment

Several features that have been developed and implemented for enhancing effects of matrix visualization for interval type symbolic data are introduced in this section. All related modules with features and methods proposed in this project are integrated into a Java version software called *i*GAP that can be freely downloaded for academic use.

4.1. Five interval data displaying modes

An interval can be converted to four numeric values: minimum, maximum, midpoint, and length. With a mouse click on the *i*GAP MV display, Fig. 8(a)–(d) displays the Minryoku 2010 data table $I_{151\times 58}$ from Fig. 7 using these four with the original

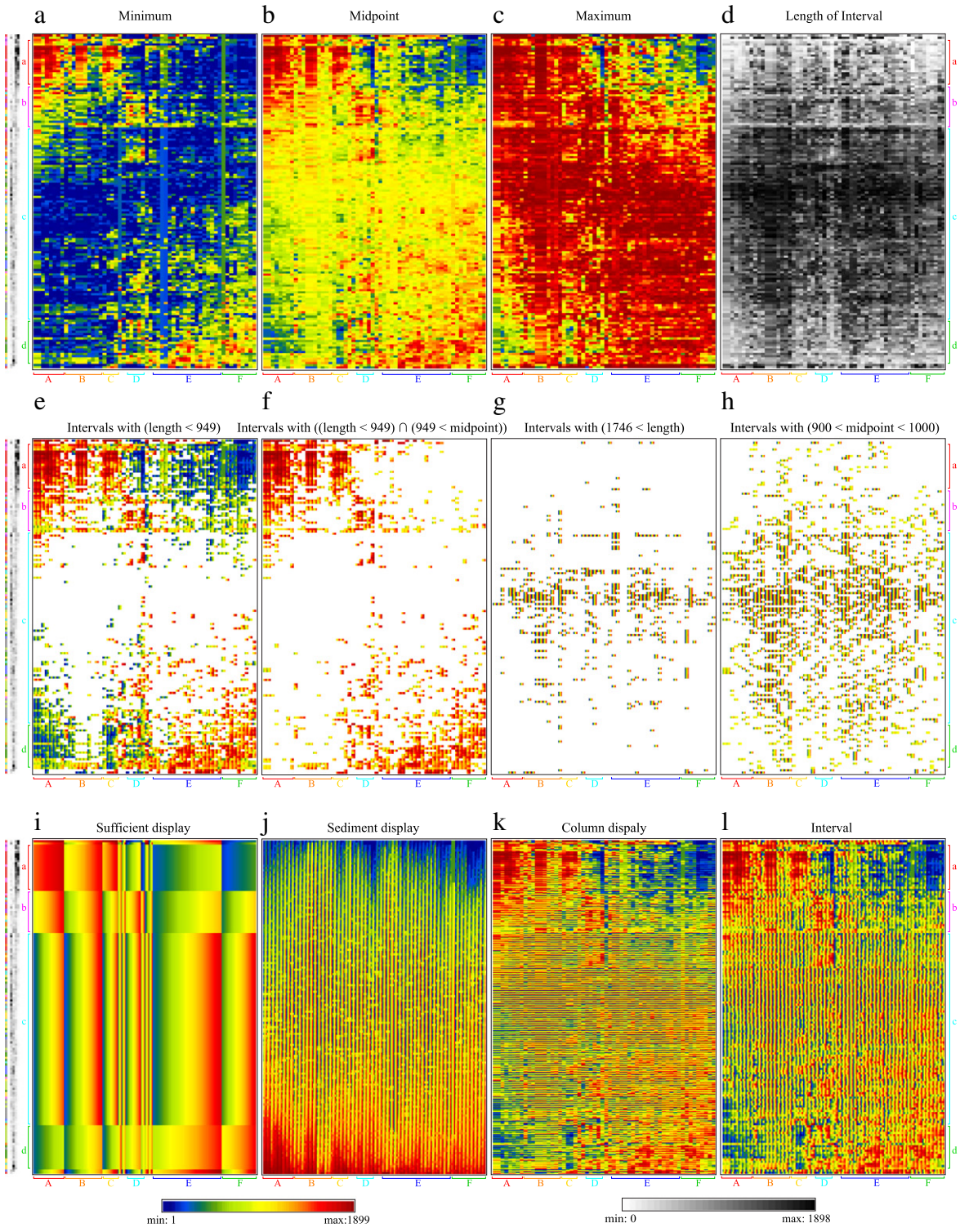

Fig. 8. Twelve displaying modes for MV of interval data for the Minryoku 2010 example in Fig. 7. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

interval displaying mode copied as Fig. 8(l) for easier comparison. By moving between these five modes of interval tables, users can pick out different aspects of the data structure. One sees that minimum values for combinations (ABC by a) and (EF by d) are still rather large (in red) in Fig. 8(a). Several vertical strips in Fig. 8(a) with identical colors in light blue and green

reveal that many regions share identical minimum values for these interval variables. The midpoint matrix map in Fig. 8(b) resembles the full interval pattern in Fig. 8(l). The maximum map in Fig. 8(c) has an opposite structure to the minimum map of Fig. 8(a). Fig. 8(d) illustrates well the length pattern of the 151 by 58 intervals. Scale for lengths of intervals usually has a rather distinct range from that of the other displaying modes and requires a different color spectrum.

4.2. Midpoint condition range display

The GAP approach MV display can visualize interval data tables with thousands of concepts on thousands of interval variables simultaneously, but it can be a tedious job to search for meaningful patterns. Users have the option of displaying only intervals that satisfy certain criteria, specifying criteria interactively through a sliding bar on the active color spectrum.

One criterion is the midpoint range, users specify a certain range of midpoint to display only those intervals with midpoints lying within that range. The left panel in Fig. 6(b) further explores the Meteorological station data from Fig. 6(a) using the midpoint criterion with the range set at (9–25 °C). With another mouse click, the right panel in Fig. 6(b) displays only those temperature intervals with midpoints colder than 9 °C or warmer than 25 °C. One sees that only three stations retain all intervals in the left panel of Fig. 6(b): KunMing, TengChong, and XiChang formed a group of outliers that disrupted the smooth pattern from north (coldest) to south (warmest). There is a Chinese saying for the city of KunMing: It is like spring all the year round. In the right panel of Fig. 6(b), we see the other 57 stations suffered either from too cold a winter or too hot a summer or both. Although KunMing, TengChong, and XiChang share the temperature range (9–25 °C) across all 12 months, XiChang has relatively warmer summer months (darker red right ends of the summer month intervals).

4.3. Length condition range display

A second display criterion is the length, users specify a certain range of lengths for displaying only those intervals with lengths satisfying that range. Fig. 8(e) further explores the Minryoku 2010 data from Fig. 8(l) using length <949 (half of the number of Level 4 cities).

Users can combine these two criteria for a more flexible interactive exploration of the interval data. As illustrated in Fig. 8(f), length <949 and midpoint >949. Fig. 8(g) displays only intervals with length >1746. From the covariate of count of cities within an area, one sees that most of the remaining regions share relatively wider ranges because they contain relatively more level 4 cities.

4.4. Sufficient, sediment, and row-condition displays

With a properly sorted and partitioned MV for conventional data, Chen (2002) introduced the concept of a sufficient graph. One replaces each partition by a summary statistic such as mean or median. A sufficient display of a sorted and partitioned interval data table identifies the summary interval for each of the partitions (intersection of concept-cluster by variable-group). Fig. 8(i) retains the summary (sufficient) pattern of all 8758 individual intervals (151 areas by 58 variables) in Table 2 with 24 summary intervals (4 area clusters by 6 variable groups) and several minor intervals for intersections regarding outlying areas and variables.

Another useful feature is the sediment display constructed by individually sorting rows within each column according to the ascending (descending) order of numerical magnitudes. Links for row profiles are lost and the interpretation is analogous to a side-by-side box-plot. For interval data, one can order intervals using minimum, midpoint, or maximum. Displayed in Fig. 8(j) is the sediment display with each variable individually sorted using midpoints of intervals for that variable.

Color strips representing intervals have been displayed horizontally (left to right for minimum to maximum) for easier vertical comparison among concepts. This is termed the column-condition display. Users can also use a row-condition display for performing horizontal comparisons (top to bottom for minimum to maximum) across variables within each concept, such as in Fig. 8(k) for Minryoku data. It is also possible to perform a row-wise sediment display within a row-condition display.

5. Discussions and concluding remarks

According to John Tukey in his EDA (1977): *It is important to understand what you CAN DO before you learn to measure how WELL you seem to have DONE it*. The SDA literature provides useful tools and algorithms for handling various types of symbolic data. Some of them act as black boxes in that users submit input data tables and obtain output results without an understanding of the mechanism involved. Our methods provide users information on the grouping of interval variables ($P^I_{p\times p}$), clustering of interval concepts ($P^C_{k\times k}$), and their interactions on the interval data table itself ($I_{k\times p}$). We hope these provide users with a detailed understanding of their interval data structure so as to understand what they CAN DO before standard assumptions and statistical procedures for symbolic data are applied, and with a way of visualizing the results of more formal procedures.

In many occasions a conventional statistical procedure for analyzing the midpoint-matrix only (instead of the interval-table) can already serve the needs. The midpoint condition range display in Section 4.2 allows users to verify if this is the case. Fig. 8(h) displays only intervals in the sorted Minryoku 2010 data table with 900 $<$ midpoints $<$ 1000. Various groups

with different lengths of interval can be identified in this map implying a reduced midpoint-only analysis may not suffice in this particular situation.

We believe SDA has the potential for handling data of a hierarchical or dependent nature, and for approaching large data. The proposed methods/techniques, with software environment for analyzing interval type symbolic data, provide a platform for exploration of the world of symbolic data. The *i*GAP environment can visualize interval data with up to 8000 variables and concepts simultaneously on a 64-bit personal computer with 8 GB of RAM. The zooming function provides reasonable readability for both detailed information and global structure for thousands of variables/concepts in a single matrix visualization display.

The current study has focused only on interval data. The developed methods/techniques can be easily adapted to such commonly used data types as continuous or ordinal multi-valued, modal multi-valued or interval-valued (histogram). The details will be described elsewhere. Many uses, in particular the interactive ones, of the methods/techniques implemented in the software *i*GAP are evident only when practiced. There are several demo datasets that allow this, available for download.

Acknowledgments

*i*GAP is available to readers and is free to non-commercial applications. The installation instructions, the user's manual, and the detailed tutorials can be found at http://gap.stat.sinica.edu.tw/Software/GAP. The authors thank two anonymous reviewers and editors for valuable and insightful comments with suggestions. They are grateful to Donald Ylvisaker and Jaromír Antoch for many valuable suggestions. The support of The Institute of Statistical Mathematics, Tokyo, Japan and the friendship of colleagues during C.H. Chen's visit in February and March, 2011 are gratefully acknowledged. This work was supported partially by the National Science Council of Taiwan, ROC under grants NSC-101-2118-M-001-007, NSC-102-2118-M-00l-011, and the National Core Facility Program for Biotechnology, Taiwan (Bioinformatics Consortium of Taiwan, NSC100-2319-B-010-002).

References

Bertin, J., 1967. Semiologie Graphique. Editions Gauthier-Villars, Paris;
English translation by Berg, William J., 1983. Semiology of Graphics: Diagrams, Networks, Maps. The University of Wisconsin Press, Madison, WI.
Bertrand, P., Diday, E., 1985. A visual representation of the compatibility between an order and a dissimilarity index: the pyramids. Comput. Stat. Q. 2 (1), 31–42.
Billard, L., Diday, E., 2000. Regression analysis for interval-valued data. In: Kiers, H.A.L., Rasson, J.-P., Groenen, P.J.F., Schader, M. (Eds.), Data Analysis, Classification, and Related Methods. Springer-Verlag, Berlin, pp. 369–374.
Billard, L., Diday, E., 2003. From the statistics of data to the statistics of knowledge: symbolic data analysis. J. Amer. Statist. Assoc. 98, 470–487.
Billard, L., Diday, E., 2006. Symbolic Data Analysis: Conceptual Statistics and Data Mining. John Wiley & Sons Ltd., England, pp. 231–248.
Billard, L., Douzal-Chouakria, A., Diday, E., 2009. Symbolic principal component for interval-valued observations, http://hal.archives-ouvertes.fr/docs/00/36/10/53/PDF/DouzalPCA.pdf.
Bock, H.-H., 2002. Clustering methods and Kohonen maps for symbolic data. J. Japanese Soc. Comput. Statist. 15, 1–13.
Bock, H.-H., 2008. Visualizing symbolic data by Kohonen maps. In: Diday, E., Noirhomme, M. (Eds.), Symbolic Data Analysis and the SODAS Software. Wiley, Chichester, pp. 205–234.
Bock, H.-H., Diday, E. (Eds.), 2000. Analysis of Symbolic Data. Springer-Verlag, Berlin, New York.
Borland, D., Taylor II, R.M., 2007. Rainbow color map (still) considered harmful. IEEE Comput. Graph. Appl. 27 (2), 14–17.
Brewer, C.A., 1994. Color use guidelines for mapping and visualization. In: MacEachren, A.M., Taylor, D.R.F. (Eds.), Visualization in Modern Cartography. Elsevier Science, Tarrytown, NY, pp. 123–147 (Chapter 7).
Brewer, C.A., 1999. Color use guidelines for data representation. In: Proceedings of the Section on Statistical Graphics. American Statistical Association, pp. 50–60.
Brito, P., 2002. Hierarchical and pyramidal clustering for symbolic data. J. Japanese Soc. Comput. Statist. 15 (2), 231–244.
Brito, P., Duarte Silva, A.P., 2012. Modelling interval data with normal and skew-normal distributions. J. Appl. Stat. 39 (1), 3–20.
Chavent, M., de Carvalho, F.A.T., Lechevallier, Y., Verde, R., 2006. New clustering methods for interval data. Comput. Statist. 21, 211–230.
Chavent, M., Lechevallier, Y., 2002. Dynamical clustering of interval data. Optimization of an adequacy criterion based on hausdorff distance. In: Jajuga, K., Sokolowski, A., Bock, H.-H. (Eds.), Classification, Clustering, and Data Analysis. Springer-Verlag, Berlin, pp. 53–59.
Chen, C.H., 2002. Generalized association plots: information visualization via iteratively generated correlation matrices. Statist. Sinica 12, 7–29.
Chen, C.H., Hwu, H.G., Jang, W.J., Kao, C.H., Tien, Y.J., Tzeng, S., Wu, H.M., 2004. Matrix visualization and information mining. In: Proceedings in Computational Statistics 2004. Compstat 2004. Physica-Verlag, Heidelberg, pp. 85–100.
Chouakria, A., Cazes, P., Diday, E., 2000. Symbolic principal component analysis. In: Bock, H.-H., Diday, E. (Eds.), Analysis of Symbolic Data. Springer, Heidelberg, pp. 200–212.
de Carvalho, F.A.T., Brito, B., Bock, H.-H., 2006. Dynamic clustering for interval data based on L_2 distance. Comput. Statist. 21, 231–250.
de Falguerolles, A., Friedrich, F., Sawitzki, G., 1997. A tribute to J. Bertins graphical data analysis. In: Bandilla, W., Faulbaum, F. (Eds.), SoftStat 97. In: Advances in Statistical Software6, Lucius & Lucius, pp. 11–20.
Denáux, T., Masson, M., 2000. Multidimensinal scaling of interval-valued dissimilarity data. Pattern Recognit. Lett. 21, 83–92.
Diday, E., 1971. La méthode des nuées dynamiques. Rev. Statist. Appl. 19 (2), 19–34.
Diday, E., 1987. The symbolic approach in clustering and related methods of data analysis. In: Bock, H.-H. (Ed.), Classification and Related Methods of Data Analysis. North-Holland, Amsterdam, pp. 673–684.
Diday, E., 2002. An introduction to symbolic data analysis and the SODAS software. J. Symb. Data Anal. 1.
Diday, E., Noirhomme-Fraiture, M. (Eds.), 2008. Symbolic Data Analysis and The SODAS Software. John Wiley & Sons Ltd., Chichester, England.
Duarte Silva, A.P., Brito, P., 2006. Linear discriminant analysis for interval data. Comput. Statist. 21 (2), 289–308.
Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D., 1998. Cluster analysis and display of genome-wide expression patterns. Proc. Natl Acad. Sci. USA 95, 14863–14868.
El Golli, A., Conan-Guez, B., Rossi, F., 2004. A self-organizing map for dissimilarity data. In: Banks, D., House, L., McMorris, F.R., Arabie, P., Gaul, W. (Eds.), Classification, Clustering, and Data Mining Applications. In: Studies in Classification, Data Analysis, and Knowledge Organization, Springer, Heidelberg, pp. 61–68.
Elmqvist, N., Do, T.-N., Goodell, H., Henry, N., Fekete, J.-D., 2008. ZAME: interactive large-scale graph visualization. In Proceedings of the IEEE Pacific Visualization Symposium, pp. 215–222.

Elmqvist, N., Dragicevic, P., Fekete, J.-D., 2011. Color lens: adaptive color scale optimization for visual exploration. IEEE Trans. Vis. Comput. Graphics 17 (6), 795–807.
Friendly, M., 2002. Corrgrams: exploratory displays for correlation matrices. Amer. Statist. 56 (4), 316–324.
Ghoniem, M., Fekete, J., Castagliola, P., 2005. On the readability of graphs using node-link and matrix-based representations: a controlled experiment and statistical analysis. Inf. Vis. 4 (2), 114–135.
Gioia, F., Lauro, N.C., 2006. Principal component analysis on interval data. Comput. Statist. 21 (2), 343–363.
Gowda, K.C., Diday, E., 1991. Symbolic clustering using a new dissimilarity measure. Pattern Recognit. 24, 567–578.
Groenen, P.J.F., Winsberg, S., Rodriguez, O., Diday, E., 2006. I-Scal: multidimensionl scaling of interval dissimilarities. Comput. Statist. Data Anal. 51, 360–378.
Guo, J., Li, W., Li, C., Gao, S., 2012. Standardization of interval symbolic data based on the empirical descriptive statistics. Comput. Statist. Data Anal. 56 (3), 602–610.
Hamada, A., Minami, H., Mizuta, M., 2008. Principal component analysis for modal interval-valued data. In: Proceedings of IASC2008, the Joint Meeting of 4th World Conference of the IASC and 6th Conference of the Asian Regional Section of the IASC on Computational Statistics & Data Analysis. ISBN 978-4-9904445-1-8, pp. 512–519.
Henry, N., Fekete, J.D., 2006. MatrixExplorer: a dual-representation system to explore social networks. IEEE Trans. Vis. Comput. Graphics 12 (5), 677–684.
Hubert, L., Arabie, P., 1985. Comparing partitions. J. Classification 2, 193–218.
Ichino, M., 1988. General metrics for mixed features-the cartesian space theory for pattern recognition. In: Proceedings of the 1988 Conference on Systems, Man, and Cybernetics. Pergamon, Oxford, pp. 494–497.
Irpino, A., Verde, R., Lauro, N.C., 2003. Visualizing symbolic data by closed shapes. In: Shader, M., Gaul, W., Vichi, M. (Eds.), Between Data Science and Applied Data Analysis. Springer-Verlag, Berlin, pp. 244–251.
Lauro, N.C., Palumbo, F., 2003. New graphical symbolic objects representations in parallel coordinates. In: Schader, Gaul, Vichi (Eds.), Between Data Science and Applied Data Analysis. Springer Verlag, pp. 288–295.
Lauro, N.C., Verde, R., Palumbo, F., 2000. Factorial discriminant analysis on symbolic objects. In: Bock, H.-H., Diday, E. (Eds.), Analysis of Symbolic Data. Springer-Verlag, Berlin, pp. 212–233.
Liiv, I., 2010. Seriation and matrix reordering methods: an historical overview. Stat. Anal. Data Min. http://dx.doi.org/10.1002/sam.10071.
Liiv, I., Opik, R., Ubi, J., Stasko, J., 2012. Visual matrix explorer for collaborative seriation. Wiley Interdiscip. Rev. Comput. Stat. 4 (1), 85–97. http://dx.doi.org/10.1002/wics.193.
Lima Neto, E.A., De Carvalho, F.A.T., 2008. Centre and range method for fitting a linear regression model to symbolic intervalar data. Comput. Statist. Data Anal. 52, 1500–1515.
Lima Neto, E.A., De Carvalho, F.A.T., 2010. Constrained linear regression models for symbolic interval-valued variables. Comput. Statist. Data Anal. 54, 333–347.
Marchette, D.J., Solka, J.L., 2003. Using data images for outlier detection. Comput. Statist. Data Anal. 43, 541–552.
Micallef, L., Dragicevic, P., Fekete, J-D., 2012. Assessing the effect of visualizations on Bayesian reasoning through crowdsourcing. IEEE Trans. Vis. Comput. Graphics 18 (12), 2536–2545.
Minami, H., Mizuta, M., 2008. Symbolic multidimensional scaling and its application for internet traffic data COMPSTAT2008. In: Porto COMPSTAT'2008 Book of Abstracts, Faculdade de Economia da Universidade do Porto, FEP, 171.
Minnotte, M., West, W., 1998. The data image: a tool for exploring high dimensional data sets. In: Proceedings of the ASA Section on Statistical Graphics, Dallas, Texas, pp. 25–33.
Noirhomme-Fraiture, M., Rouard, M., 2000. Visualizing and editing symbolic objects. In: Bock, H.-H., Diday, E. (Eds.), Analysis of Symbolic Data. Springer-Verlag, Berlin, pp. 125–138.
Palumbo, F., Lauro, N.C., 2003. A PCA for interval valued data based on midpoints and radii. In: Yanai, H., Okada, A., Shigemasu, K., Kano, Y., Meulman, J.J. (Eds.), New Developments in Psychometrics. Springer-Verlag, Tokyo, pp. 641–648.
Roger, D.P., 2008. A method for visualizing multivariate time series data. J. Stat. Softw. 25 (1), 1–17.
Rosenberg, N.A., 2003. DISTRUCT: a program for the graphical display of population structure. Mol. Ecol. Notes 4 (1), 137–138.
Saito, T., Miyamura, H.N., Yamamoto, M., Saito, H., Hoshiya, Y., Kaseda, T., 2005. Two-tone pseudo coloring: compact visualization for one-dimensional data. In: Proceedings of the IEEE Symposium on Information Visualization, pp. 173–180.
Sokal, R.R., Rohlf, F.J., 1962. The comparison of dendrograms by objective methods. Taxon 11, 33–40.
Souza, R.M.C.R., de Carvalho, F.A.T., 2004. Clustering of interval data based on city-block distances. Pattern Recognit. Lett. 25 (3), 353–365.
Tien, Y.J., Lee, Y.S, Wu, H.M., Chen, C.H., 2008. Methods for simultaneously identifying coherent local clusters with smooth global patterns in gene expression profiles. BMC Bioinformatics 9, 155.
Tukey, J.W., 1977. Exploratory Data Analysis. Addison-Wesley.
Verde, R., Lechevallier, Y., 2005. Crossed clustering method on symbolic data tables. In: New Developments in Classification and Data Analysis. Springer, pp. 87–94.
Verde, R., Lechevallier, Y., Chavent, M., 2003. Symbolic clustering interpretation and visualization. J. Symb. Data Anal. 1 (1).
Ware, C., 2004. Information Visualization: Perception for Design, second ed. Morgan Kaufmann, pp. 103–149.
Wegman, E.J., 1990. Hyperdimensional data analysis using parallel coordinates. J. Amer. Statist. Assoc. 85 (411), 664–675.
Weinstein, J.N., 2008. A postgenomic visual icon. Science 319, 1772–1773.
Wijffelaars, M., Vliegen, R., van Wijk, J., 2008. Generating color palettes using intuitive parameters. Comput. Graph. Forum 27 (3), 743–750.
Wilkinson, L., Friendly, M., 2009. The history of the cluster heat map. Amer. Statist. 63 (2), 179–184.
Wu, H.M., Tien, Y.J., Chen, C.H., 2010. GAP: a graphical environment for matrix visualization and cluster analysis. Comput. Statist. Data Anal. 54, 767–778.
Wu, H.M., Tzeng, S., Chen, C.H., 2008. Matrix visualization. In: Chen, Chun-houh, Hardle, Wolfgang, Unwin, Antony (Eds.), Handbook of Computational Statistics (Volume III): Data Visualization. Springer-Verlag, Heidelberg.