The Cluster Analyzer & Visualizer

—

a Prototype for Data Exploration

Johan Hagman

SSKKII / Dept of Linguistics
University of Göteborg
hagman@ling.gu.se

SSKKII Report #5

SSKKII
Center for Studies of
Language, Cognition, Interaction,
Semantics, Communication, and Information

Abstract

The paper shows how two powerful techniques for supporting data exploration of multidimensional data have been combined in a tool for this purpose. The techniques, cluster analysis *and graphic* visualization, *are briefly presented and discussed as modules of the prototype. Its performance is illustrated by experiments resulting in cluster configurations where the value distribution of "underlying" dimensions is visualized with colours or shades of grey. Other reported experiments concern the validation of obtained cluster configurations; comparisons between clustering+visualization and neural networks; and variants of continuous isosurface colouring.*

Keywords

information exploration, visualization, cluster analysis, multidimensional analysis, isosurface colouring / shading, neural networks, cluster validation, projections, maps, correlations, human-computer interaction

Contents

1 Introduction 1
|
2 The Structure of CAV 1
2.1 Data Import Format |
2.2 Data processing |
2.2.1 Pattern matrix |
2.2.2 Proximity matrix |
2.2.3 Dendrogram 2
2.2.4 Iterative procedure |
2.3 Visualization 3
|
3 Experiments 4
3.1 Regeneration of a Map from a Road Distance Table 5
3.2 Cluster Analysis and Neural Networks 6
— a Comparison |
3.3 Colouring and shading |
|
4 Conclusion and Further Work 8
|
Three colour plates 9
|
References 12

1 Introduction

A system which integrates *cluster analysis* and *visualization* has been developed to meet an often felt need for this special kind of tool for data exploration in scientific, industrial and everyday applications. The prototype has preliminary been called the *Cluster Analyzer & Visualizer* — or *CAV* for short and this paper discusses some experiences from its development and present performance.

The epoch we are living in at present is getting less "industrial" and more "informational". In some areas this process has been going on to such an extent that we are presented with difficulties in surveying all the information offered and we find it hard to navigate through it to get what we want or to discover unexpected things. The tool presented here is an attempt to combine two techniques of making especially multidimensional data more graspable — *cluster analysis* and *data visualization*. The development of these techniques in the latest 25 years or so has essentially been parallel with that of the computers and their usability depends heavily on the latter's performance.

2 The Structure of CAV

The development of CAV has included several aspects:

- requirements presented by a number of "real-world" applications
- choice of clustering methodology, algorithms, and implementation environment
- experiments with various types of visualization
- steps toward an integration with human-computer interactive facilities, aiming at a powerful tool for information exploration and visualization.

In [Hagman 94] these aspects are presented more in detail along with the terminology that will be used here.

2.1 Data import format

CAV accepts *pattern matrices* (e.g. TABLE 2.1) and *proximity matrices* (TABLE 2.3) in plain ASCII format either as tables with rows and columns or as lists (in case the table would be too big to handle with available editors). Empty cells are allowed for and the proximity matrices are supposed to be marked as to whether the values denote *similarity* or *dissimilarity*.

2.2 Data processing

In the current version the data are equalized *before* they are normalized. Since equalization works with *relative* values such as the proportion of the standard (or maximal) deviation to the mean value for each feature[1], the values do not have to be within the same range when comparing their *relative* spread. CAV checks both the mean deviation and the maximal deviation in order to detect single, extreme cases. Then, when the internal variation has been "compressed" or "blowed up" to be within a user-specified relative range, the values are normalized.

Missing data do not affect the equalization. Mean and maximum values can be calculated anyway out of the data actually supplied. One suggestion is e.g. that missing data can be filled in by the mean value. Another, more elaborate proposal is that a pattern which is lacking a value for a certain feature could have this value calculated by comparing (seemingly ignoring other possibly absent data) with the *n* patterns most similar to that pattern.

2.2.1 Pattern matrix

The operations on the original matrix result in a normalized and equalized pattern matrix. The output of all three steps described in 2.2.1 through 2.2.3 may be downloaded to separate files for special examination. Below, in section 2.3, yet another method is proposed.

2.2.2 Proximity matrix

The pattern matrix underlies the calculation of the proximity matrix. For each pair of patterns the sum or product[2] of the differences of their corresponding feature values is calculated.

---

[1] Or 'attribute', 'parameter', or 'column'.

[2] Depending on whether *Manhattan distance* or *Euclidean distance* is preferred.

Sums are then divided by the number of comparable features (i.e. where neither pattern had an empty cell), giving the proximity index for that pattern pair. Eventually this process yields a proximity matrix that could be used directly as input for future runs. It is also possible to analyze the pattern matrix vertically instead of horizontally. Then, instead, the *features* are compared as to their covariance through the patterns and the subsequent cluster configurations will show clustering tendency of these features. Whatever is analyzed, patterns or features, the term 'item' will be used henceforth.

2.2.3 Dendrogram

The proximity matrix is passed on to a procedure which builds the corresponding binary dendrogram (e.g. FIGURE 3.3a, below). At present, an agglomerative, hierarchical clustering is carried out which iteratively merges the rows and columns of the most similar pair of items while simultaneously joining the corresponding subtrees in a new node.

2.2.4 Iterative procedure

The next phase is an "animated" clustering, currently in a 2D universe. The algorithm is a variant of the iterative one presented in [Sammon 69]. In CAV the items are represented by small, mobile labels on a 2D plane where they move around, driven by their interrelational forces of attraction and repulsion, like the balls in billiards (FIGURE 2.4). They continue to move until a satisfactory low-stress configuration has been reached. The criterion for satisfaction is given by the user; a typical one is to let the configuration adjust its shape until no new minimal stress value has been encountered during the latest *n* movements, having *n* set to some suitable value. Different initial configurations have been tried for the same proximity matrix and with very rare exceptions they yield the same final result. The crucial thing is to avoid cases where items are too astray to make it on time to reach their "home cluster" before other clusters are formed in its way blocking it. To minimize this risk the dendrogram is used; the order of its leaves from left to right is a kind of 1D representation of their internal order and that is followed when the items are lined up diagonally over the screen[3]. This speeds up the 2D clustering process considerably and straightens the item trajectories.

PM	1	2	3	4	5	6	7	8
'8'	7	13	5	5	6	13	2	3
'8'	5	13	6	4	6	13	3	13
'8'	9	10	6	6	8	10	2	3
'8'	7	7	6	6	8	7	2	3
'8'	8	7	6	6	8	7	2	0
'8'	7	7	6	7	7	7	1	1
'8'	6	10	7	8	8	9	4	4
'8'	6	7	7	7	9	8	4	5
'8'	5	5	5	12	10	7	2	3
'8'	7	8	4	4	7	6	2	3
'8'	8	7	5	4	6	10	1	0
'8'	6	10	5	2	6	8	1	2
'8'	7	10	5	5	8	7	1	20
'8'	7	10	6	6	6	8	3	3
'8'	6	6	7	7	8	8	3	2
'O'	7	7	5	6	3	3	4	6
'O'	7	6	7	6	3	4	6	5
'O'	6	6	5	5	4	3	4	5
'O'	8	8	7	6	5	7	5	5
'O'	6	8	7	5	5	6	2	2
'O'	7	7	7	7	6	6	3	2
'O'	7	7	7	7	5	6	5	5
'O'	7	7	8	7	5	6	4	5
'O'	8	7	7	6	6	5	4	5
'O'	7	8	6	5	4	6	2	4
'O'	9	7	7	6	8	6	4	3
'O'	8	8	7	6	7	6	4	4
'O'	7	6	6	5	3	2	5	4
'O'	7	6	7	5	5	5	4	3
'O'	0	8	7	6	6	7	2	4
'X'	10	7	6	6	9	9	7	10
'X'	10	4	4	4	8	8	3	10
'X'	10	7	4	4	9	9	3	9
'X'	7	7	6	5	10	10	10	8
'X'	6	10	6	10	8	8	13	4
'X'	8	10	7	10	9	8	11	4
'X'	5	7	8	7	8	9	9	8
'X'	5	6	7	7	9	9	9	7
'X'	6	7	11	6	8	11	7	10
'X'	6	12	10	4	8	11	8	3
'X'	7	12	8	6	9	11	9	1
'X'	11	8	7	10	11	10	6	9
'X'	9	5	6	7	10	9	7	5
'X'	10	5	6	4	9	9	6	11
'X'	0	5	11	6	9	11	5	9

TABLE 1a Pattern matrix for the '8OX' data; eight features describe fifteen variants of each of the three characters '8', 'O', and 'X'

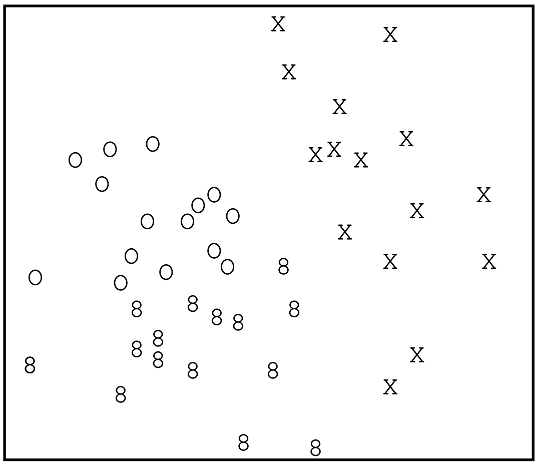


FIGURE 1b Euclidean projection of the '8OX' data by [Sammon 69] from [Jain& Dubes 88].

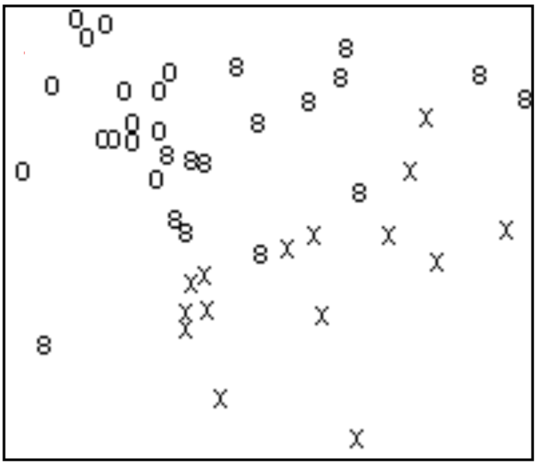


FIGURE 1c Manhattan projection of the '8OX' data by CAV.

A comparison of [Sammon 69]'s Euclidean-based projection [Jain & Dubes 88, p. 39] of the so-called '8OX data' has also been made with a CAV Minkowski-based projection (FIGURES 1b and 1c). The iterative procedure gives good results and today's computers, as compared with those of the '60s, offer possibilities of another magnitude concerning execution time and memory. Other methods that also involve an iterative reshaping of the configuration have been proposed; one of them is *simulated annealing* [Kirkpatrick 83], [Bell 90], [Chalmers & Chitson 92] and below *neural networks* represent yet another.

---

[3] An alternative is to start by forming a circle.

When the best configuration has been found, a validation of the clustering can be made. In a series of tests the program was fed with pattern matrixes or proximity matrices describing 2D positions of corners of various known geometrical shapes which later during the clustering process have gradually re-emerged on the display. One of these experiments is illustrated by TABLES 2.1 and 2.3, and FIGURES 2.2 and 2.4, below. Another example is found in section 3.1.

12 2	Feature1	Feature2
'1.1'	1	1
'1.2'	1	2
'1.3'	1	3
'1.4'	1	4
'1.5'	1	5
'2.3'	2	3
'3.1'	3	1
'3.3'	3	3
'3.5'	3	5
'5.1'	5	1
'5.3'	5	3
'5.5'	5	5

TABLE 2.1 Pattern matrix with 12 patterns, each of which is described with two features and named after these values. Item '1.3' has been encircled.

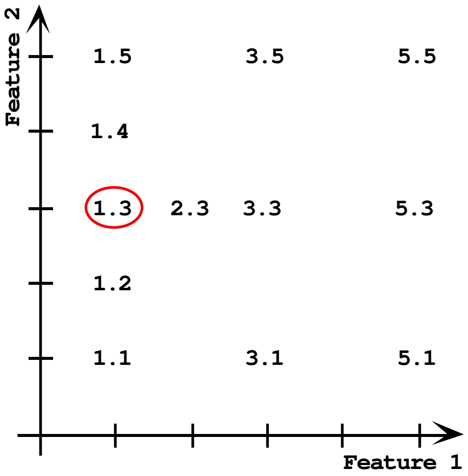


FIGURE 2.2 An intuitive 2D projection of the simple matrix in TABLE 2.1. Item '1.3' has been encircled.

D 12	'1.1'	'1.2'	'1.3'	'1.4'	'1.5'	'2.3'	'3.1'	'3.3'	'3.5'	'5.1'	'5.3'	'5.5'
'1.1'	–	1	2	3	4	3	2	4	6	4	6	8
'1.2'	1	–	1	2	3	2	3	3	5	5	5	7
'1.3'	2	1	–	1	2	1	4	2	4	6	4	6
'1.4'	3	2	1	–	1	2	5	3	3	7	5	5
'1.5'	4	3	2	1	–	3	6	4	2	8	6	4
'2.3'	3	2	1	2	3	–	3	1	3	5	3	5
'3.1'	2	3	4	5	6	3	–	2	4	2	4	6
'3.3'	4	3	2	3	4	1	2	–	2	4	2	4
'3.5'	6	5	4	3	2	3	4	2	–	6	4	2
'5.1'	4	5	6	7	8	5	2	4	6	–	2	4
'5.3'	6	5	4	5	6	3	4	2	4	2	–	2
'5.5'	8	7	6	5	4	5	6	4	2	4	2	–

TABLE 2.3 Proximity matrix showing the "Manhattan" interdistances of the 12 patterns in TABLE 2.1. The row and column relevant to item '1.3' have been indicated.

2.3 Visualization

We proceed to the next module, which is a more advanced visualization where shades of grey or colour are added to the 2D cluster configuration. Both the grey shades and the colours constitute interval scales with a finite number of degrees which, however, depending on the hardware, may be millions. Grey shades increase with the values from black to white. The colour scale also runs from black to white but passes through blue, purple, red, orange, and yellow. Note that colour intervals may be differently defined; frequently (mostly) the interval is an inverted variant of the visible spectrum with the increasing sequence purple, blue, green, yellow, orange, and red. This could be claimed to be a "natural" interval but even the choice of the aforementioned interval could be supported by nature, i.e. e.g. an analogy of metal at different temperatures (assuming a bluish metal) which explains the absence of green. Anyway, [Brodlie & al. 92] state that:

> As yet there are no widely accepted standard colour scales or range scales (such as logarithmic) for data exploration in visualization. A need is perceived for empirical studies leading to the derivation of formal guidelines and reliable standards in this area. [p. 119]

For another view of colours, see e.g. [Tufte 83].

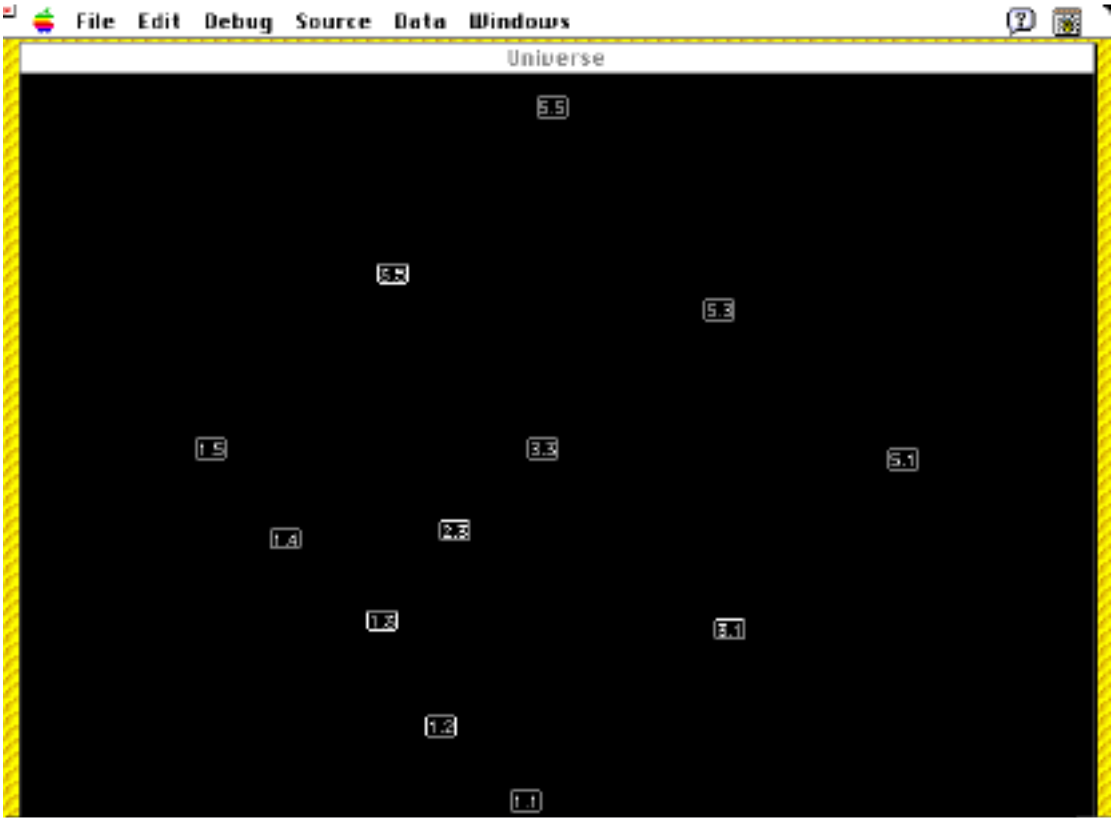


FIGURE 2.4 The final cluster configuration obtained in a *CAV* session. The twelve items attract/repel each other according to their degree of similarity shown in TABLE 2.3.

Shades or colours (henceforth: 'intensity'[4]) can be used to enhance the groupings of the clusters so that for each picture element (i.e. the square of $n*n$ pixels, $n \geq 1$) of the picture the density of items in that part of the configuration is represented by a value on the intensity interval. To do this, the minimum and maximum densities have to be calculated in order to calibrate the intensity scale and thereby use it optimally. When the impression of groups in this way gets enhanced, each item contributes to its "background" shading/colouring by its mere presence. All items contribute equally to the intensity distribution throughout the picture. This is illustrated e.g. by the FIGURE 3.4a.

Now, it is also possible to associate different values to the items — values, for instance, corresponding to one of the underlying features (if the items correspond to the *rows* of an underlying pattern matrix). In such a visualization, the intensity would reflect the distribution of that feature throughout the picture. Examples of this are the FIGURES 4.1a and 4.1c. Items with missing data for the chosen feature can have their background intensity decided by interpolating the intensities of the surrounding items. This could also be a way of *estimating* a missing value, probably better than the assignment of the mean value, mentioned above, and fully comparable with the method using the nearest neighbours.

3 Experiments

One of the many pattern matrices processed by CAV is a table[5] of the consumption per capita of three types of alcoholic beverage (spirits, wine, and beer) in 32 countries. The resulting 3D⇒2D reduction and cluster configurations are showed in FIGURES 4.1a through 4.3Σ.

---

[4] Note that this refers to the values of the underlying data features and not to the brightness on the screen.

[5] Extracted from [SCB 91]

Another pattern matrix, provided by [Hofstede 91], characterizes some fifty countries according to the four dimensions 'individuality', 'masculinity', 'power distance', and 'unceritainty avoidance'. The cluster configuration of this 4D⇒2D reduction is shown as part of the FIGURES 5.1 and 5.2.

CAV has also been fed directly with proximity matrices. "Semantic spaces" for part of the content of a Swedish thesaurus have been generated where the words were clustered according to their co-appearance in definitions. The arrival of computers has made possible many interesting experiments with automatic (re)construction of thesauri [Morris & Hirst 91], [Crouch & Yang 92]. In addition, we are currently experimenting with clustering newspaper articles of different genres based on their "categorimatic" or "syncategorimatic" profiles.

A special type of proximity matrix is the *asymmetric* one, i.e. where the relation *A*→*B* is *not* the same as the relation *B*→*A*. An example of this is the speakers' asymmetric sequencing in a discourse and such a matrix was analyzed and visualized in e.g. [Allwood & Hagman 94]. Moreover, similar experiments have been done with the same data as are partly presented in [Eikmeyer 92] (i.e. association and confusion matrixes) — in collaboration with this author. One of those experiments is the following.

3.1 Regeneration of a Map from a Road Distance Table

In this experiment a table of road distances between 11 cities (in FIGURE 3.1) was fed into CAV as a *dissimilarity* matrix. TABLE 3.2 shows this matrix and FIGURE 3.3a shows the

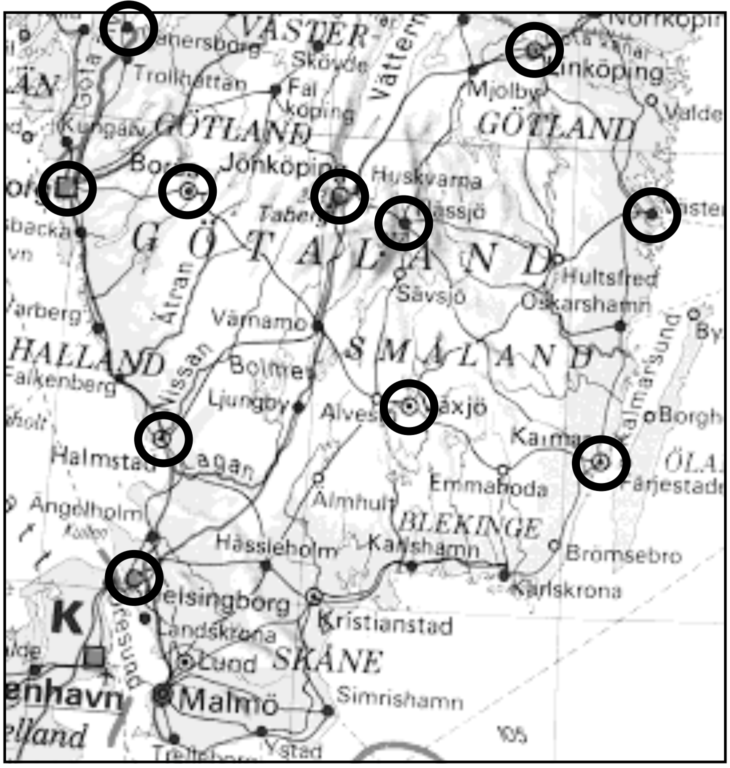

VÄSTER-
Vänersborg
Trollhättan
Skövde
Falköping
Vättern
Norrköping
Linköping
Mjölby
GÖTLAND
Valdemarsvik
Kungälv
GÖTLAND
Borås
Jönköping
Huskvarna
Taberg
Nässjö
Västervik
G Ö T A L A N D
Hultsfred
Sävsjö
Oskarshamn
Värnamo
Varberg
Ätran
HALLAND
Falkenberg
SMÅLAND
Bolmen
Nissan
Ljungby
Alvesta
Växjö
Borgholm
Halmstad
Lagan
Kalmar
ÖLAND
Färjestaden
Ängelholm
Älmhult
Emmaboda
BLEKINGE
Hässleholm
Karlshamn
Brömsebro
Karlskrona
Helsingborg
Kristianstad
Landskrona
Lund
SKÅNE
Öresund
Malmö
Simrishamn
Ystad
Trelleborg
Köbenhavn

FIGURE 3.1 (above left)
Map of southern Sweden with 11 cities encircled.

TABLE 3.2 (above right)
Road distance table (in kilometers), used as a dissimilarity matrix.

DM 11	Bo	Gö	Ha	He	Jö	Ka	Li	Nä	Vb	Vv	Vx
Borås	-	67	140	224	83	280	211	126	102	261	172
Göteborg	67	-	142	228	151	347	280	195	90	328	239
Halmstad	140	142	-	84	171	242	300	182	231	312	135
Helsingborg	224	228	84	-	242	273	370	250	316	363	184
Jönköping	83	151	171	242	-	210	129	44	149	177	122
Kalmar	280	347	242	273	210	-	227	167	359	143	107
Linköping	211	280	300	370	129	227	-	123	272	107	218
Nässjö	126	195	182	250	44	167	123	-	193	139	107
Vänersborg	102	90	231	316	149	359	272	193	-	326	271
Västervik	261	328	312	363	177	143	107	139	326	-	179
Växjö	172	239	135	184	122	107	218	107	271	179	-

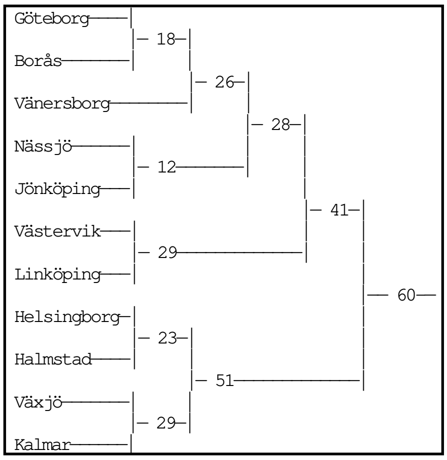

Göteborg
Borås
18
26
Vänersborg
28
Nässjö
12
Jönköping
41
Västervik
29
Linköping
60
Helsingborg
23
Halmstad
51
Växjö
29
Kalmar

FIGURE 3.3a (left)
Dendrogram reflecting city groupings. Each number indicates the distance between the two subtrees expressed as percentage of the maximum intercity distance in Figure 3.2.

resulting dendrogram after a single-linkage hierarchical clustering based on Manhattan distances. After letting the forces of attraction and repulsion move the city labels around on the 2D "animated" clustering display for a couple of minutes, the low-stress configuration shown in the final picture of FIGURE 3.4a is reached. Note however that, of course, this configuration has no *dimensions* or explicit *axes*; its form is completely determined by the intercity *relations*. The fact that even the orientation of this picture so closely coincides with FIGURE 3.1 is accidental; it could as well have been rotated or mirror-inverted.

3.2 Cluster Analysis and Neural Networks — a Comparison

The dissimilarity matrix in FIGURE 3.1 was also used as input to a self-organizing Kohonen map[6]. Output of this approach is shown in TABLES 3.3b and 3.4b. TABLE 3.3b shows the similarities between a city's distance pattern to all other cities, where the mean similarity distance is assumed to be 100 with a standard deviation of 27. A low similarity distance indicates that the cities have a highly similar distance pattern. The cities have been put together into five groups for which the group internal similarity is significantly high (i.e. the similarity distance is below 100-27=73). Compare this table with the dendrogram in FIGURE 3.3a.

	He	**Ha**	**Bo**	**Gö**	**Vb**	**Jö**	**Nä**	**Vx**	**Ka**	**Vv**	**Li**
He		**47**	104	109	117	87	90	119	126	123	121
Ha			103	108	117	90	92	123	*130*	126	124
Bo				**50**	**62**	79	87	126	*131*	*128*	*127*
Gö					**57**	85	92	*132*	*136*	*134*	*132*
Vb						77	83	120	122	120	119
Jö							**50**	82	86	83	82
Nä								77	79	79	77
Vx									**52**	106	109
Ka										98	103
Vv											**50**
Li											

FIGURE 3.3b City groupings (framed cells) as suggested by the neural network in this modified version of TABLE 3.2 .The mean distance here is 100 (Courtesy of H.-J. Eikmeyer).

The data were represented in the neural network by a 2D 10x10-cell Kohonen map and one of the best output results is shown in FIGURE 3.4b, where the city abbreviations in the cells reflect the cities' "area of dominance" and distance relations[7]. As well as in FIGURE 3.4a, even in this configuration there are no "natural axes" as could be claimed for FIGURE 3.1 but, nevertheless, the similarity among all these three figures becomes clearer when we turn FIGURE 3.4b 90° clockwise and then rotate it along its vertical Vb—Ka/He axis.

3.3 Colouring and shading

The intensity shading or colouring of the 2D configurations has invited to many experiments. The principle of assigning intensity to each point of the picture was described above. In one type of visualizations[8] the *isosurfaces* (areas of the same intensity) are indicated only by their boundaries, the *isolines*. Often height is marked in this way on topographic maps. When a geometric third dimension as height is visualized it often has to be shown in perspective to render the effect of 3D. Such "3D" visualizations could also be combined with intensity hues that either show the same feature as does the height dimension (i.e. the same colours on the same levels) or that show yet another feature with intensity fields running over hills and valleys independently of these.

All the FIGURES 4.1a through 4.3Σ are based on the same set of data, mentioned above, and their cluster configuration is identical. In FIGURES 4.1a and 4.1c the value distribution of the features 'wine consumption' and 'spirits consumption', respectively, is visualized. Note, by the way, that the items in this case already have a natural configuration suggested by their

---

[6] By H.-J. Eikmeyer and his colleagues at the Dept of Linguistics, University of Bielefeld. For related reading, see e.g. [Ritter & Kohonen 89] and [Eikmeyer 92].

[7] For an interesting analogy: cf. [Lin 92].

[8] See [Brodlie & al. 92] for a classification of visualization types.

Bo	Bo	Bo	BoGo	Go	GoHa	Ha	Ha	Ha	Ha
Bo	Bo	BoGo	Go	Go	Go	GoHa	Ha	Ha	HaHe
Vb	BoVb	BoGo	Go	Go	Go	GoHa	HaHe	He	He
Vb	Vb	NäVb	GoNä	GoNä	GoJo	Jo	HeJo	He	He
Vb	Vb	NäVb	Nä	JoNä	Jo	Jo	Jo	He	He
Vb	NäVb	Nä	Nä	Nä	Jo	Jo	Jo	JoKa	Ka
LiVb	LiNä Vb	Nä	Nä	JoNä	Jo	Jo	JoKa	Ka	Ka
Li	Li	LiNä	NäVv	NäVv	JoVv	Jo	JoKa Vx	KaVx	Ka
Li	Li	LiVv	Vv	Vv	Vv	VvVx	Vx	Vx	KaVx
Li	Li	LiVv	Vv	Vv	Vv	VvVx	Vx	Vx	Vx

FIGURE 3.4b Kohonen map based on TABLE 3.2 (H. -J. Eikmeyer)

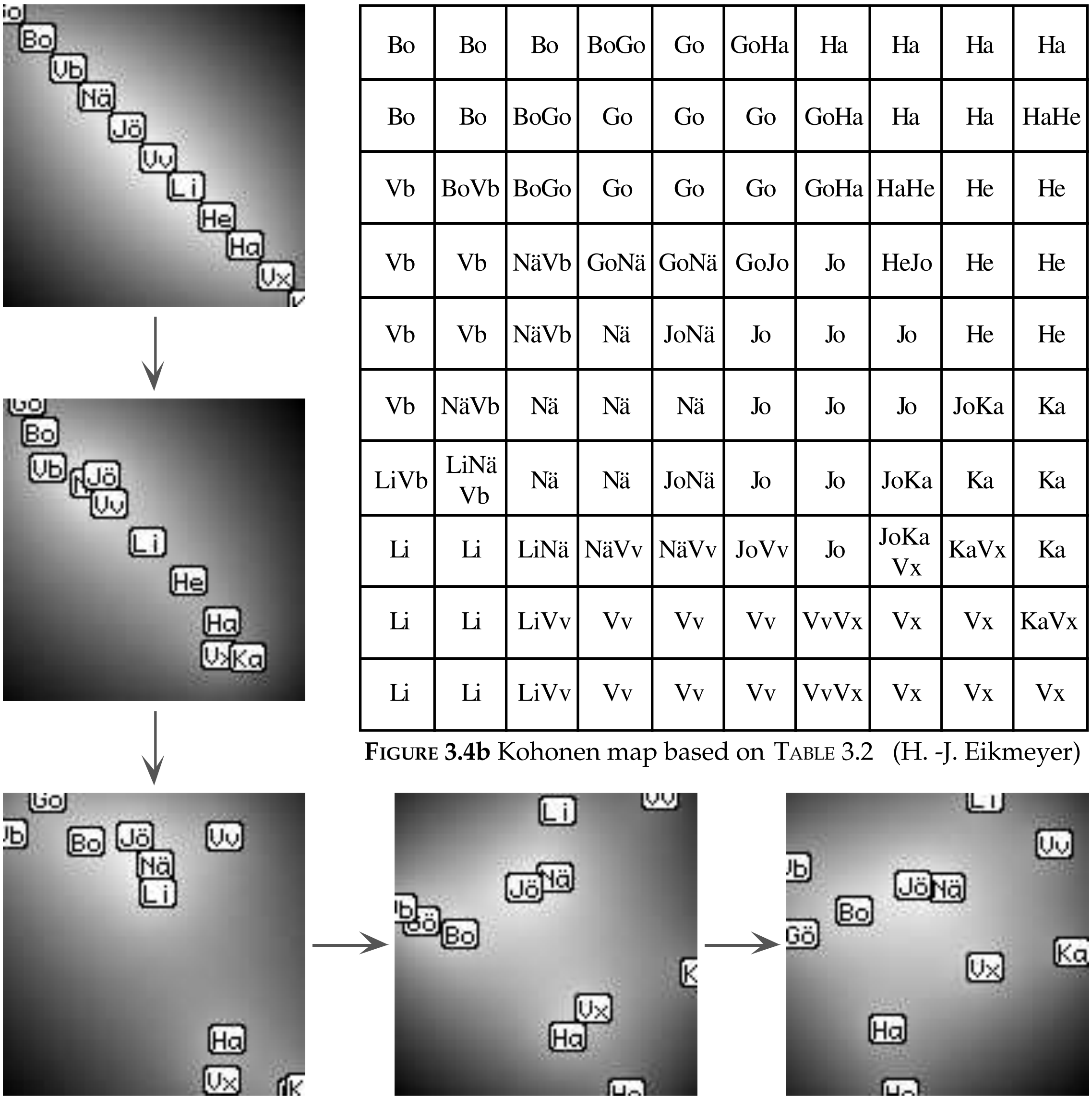


FIGURE 3.4a Stepwise reshapening in five consecutive cluster configurations generated by CAV. Starting from a diagonal line-up of the leaves of the dendrogram (FIGURE 3.3a), the 'stress' decreases as the city labels (centered at the actual city coordinates) move according to the internal forces of attraction and repulsion among the cities. The 'stress' is the discrepancy between the distances in the dissimilarity matrix (TABLE 3.2) and the current configuration.

geographical extension and that this as well could have been chosen as the configuration. Then, however, we would suffer the disadvantage of having large areas to which no values are associated (e.g. the areas between the European countries and JAPan, South AFrica, or PERu). To illustrate a situation of data exploration and the possibility to colour the configuration according to other features *which have not contributed to the clustering of the data*, FIGURES 4x and 4y are added. There, in the same configuration the reported rates of two different causes of death are visualized.

The cluster configurations of the [Hofstede 91] data in FIGURES 5.1, and 5.2 have been briefly commented upon. In Figure 5.1 the value distribution of the feature dimension 'individuality' is visualized with intensity less smoothed than it is in FIGURE 5.2 which shows the value distribution of the feature 'masculinity'.

Since all colours can be generated by combining the three elementary colours red, green, and blue, a special possibility is offered when one wants to visualize the value distribution of exactly three features on a 2D or 3D scatter plot or cluster configuration; each feature is simply associated with one of the three colours. This is described e.g. in [Ammermann & Cavalli-Sforza 84] where each colour was associated with one of the three *principal* *components* of a genetic analysis of the populations of Europe to try a hypothesis of prehistoric migration waves from Asia. The 2D representation in that study was a map over Europe and part of Asia and it was coloured in this way ...

> ... to take advantage of the power of the human eye to sum elementary colors and thus synthesize the information contained in the first three components, which together account for almost 60% of the information present in the original set of genetic data. It was heartening to find that the patterns obtained were in basic agreement with those expected under the demic hypothesis." [p. 105]

This technique has recently been applied successfully also in [Cavalli-Sforza & al. 94].

FIGURES 4.2a and 4.2c are monochromatic variants of FIGURES 4.1a and 4.1c, respectively, and FIGURES 4.3a through 4.3Σ are variants of FIGURES 4.2a through 4.2Σ with another "smoothness" of the colour shading. As opposed to FIGURES 4.1a and 4.1c, FIGURES 4.2a and 4.2c are coloured with shades of just one colour: red and blue, respectively. FIGURE 4.2b visualizes the distribution of the beer consumption per capita in the configuration with shades of green. The synthesis of three monochromatic pictures is shown in FIGURES 4.2Σ and 4.3Σ where all profiles of the consumption of 'wine', 'beer', 'liquor', and '(officially) neither of the three' are clearly visualized (not without some aesthetic appeal). Millions of colour nuances of the spectrum show the transitions from one consumption profile to the other. In FIGURE 4.2Σ the clusters of countries with a similar consumption profile continuously transcend into each other with a "smoothed-out" colour shading, wheras in FIGURE 4.3Σ they are more separated due to the dimming effect caused by that degree of smoothing.

4 Conclusion and Further Work

At present, the modules may be likened to the "turnkeys" [Brodlie & al. 92] which can be run either independently or as integrated parts of a so-called application building system. Steps are taken to integrate all or part of the modules in a more powerful system for information exploration and visualization. The intent is to fuse CAV's possibilities with those offered e.g. by *dynamic queries* and *tight-coupled starfield displays* [Ahlberg & Shneiderman 94], as well as other query devices [Ahlberg & Truvé 94]. In this integrated system, the CAV parts will contribute with the displaying of data in cluster configurations and colouring this configuration universe or "starfield" in the ways presented here. FIGURES 5.1 and 5.2 are early sketches which illustrate what a simple version of this emerging tool could look like. These figures represent computer screens where the displays' current output is a cluster configuration of data from [Hofstede 91]. The only widgets shown are the four sliders, automatically created when reading the input file and assigned to one dimension, plus a regulator for the intensity representation of selected features.

Both the kind and the degree of the *interaction* of a system for data exploration is highly depending on the application and the expected type of users. As for university or industrial researchers, they are to a greater extent expected to know their data sufficiently to try different techniques for data processing and visualization. The extreme opposite type of application is the public service tools, e.g. automatic library guides or similar [Salton & al. 94]. In such a system a subset of all possible facilities may be chosen by the system developer, leaving to the user less (but sufficient) freedom — and maybe less problems. Since the requirements of interaction are not yet presented for the integrated tool for this information exploration and visualization, its interactive part is still to be completed. Thus, its final design will depend on application, user types, and , of course, hardware environment.

Acknowledgements

This work was enabled by support from NUTEK, the Swedish National Board for Industrial and Technical Development, grant no. 5321-93-2760.

I would like to thank Hans-Jürgen Eikmeyer, Dept of Linguistics, University of Bielefeld for our fruitful collaboration and my colleagues Christopher Ahlberg, Jens Allwood, and Staffan Truvé at SSKKII Research Center (an intersection of the University of Göteborg & Chalmers University of Technology) for valuable comments and suggestions.

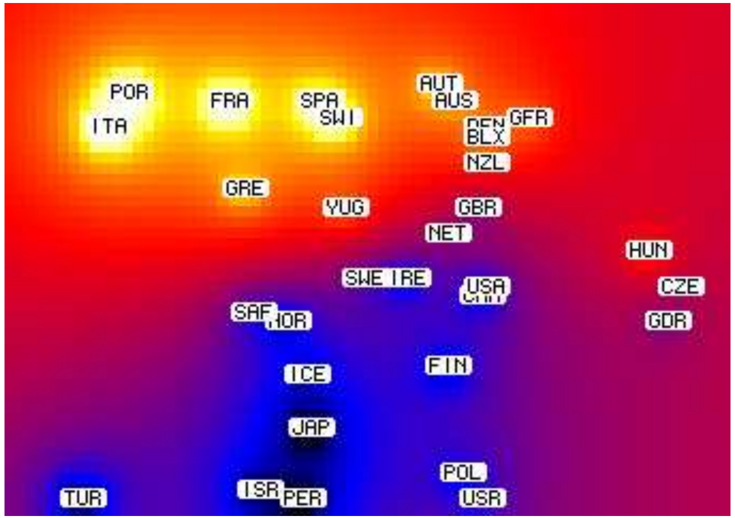


FIGURE 4.1a Cluster config. of the (reported) consumption of beer, wine, and spirits 1983-87, highlighting wine consumption.

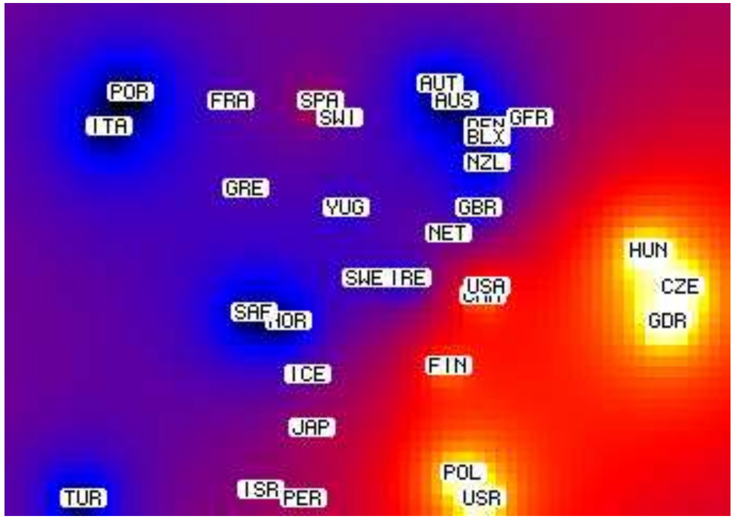


FIGURE 4.1c Same cluster as in FIGURE 4.1a. Highlights spirits consumption. (Values missing for GRE, ISR and PER)

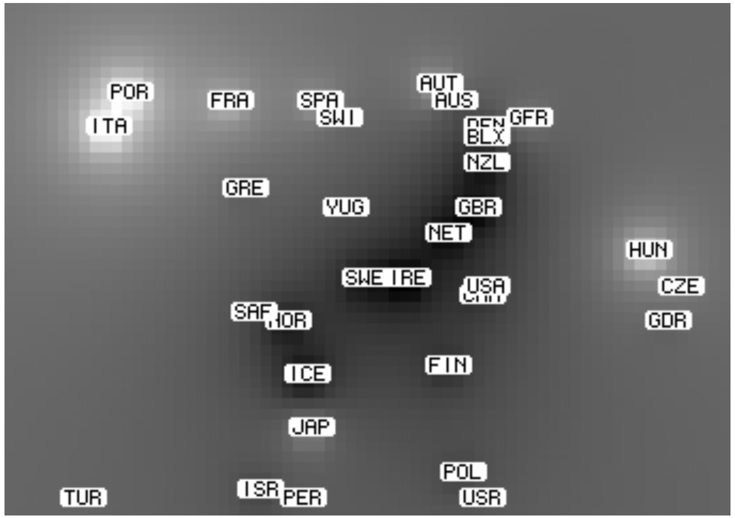


FIGURE 4x Clustered as FIGURES 4.1a & 4.1c. Deaths by cirrhosis 1982-86. (No values reported for PER,SAF,TUR,USR,and YUG)

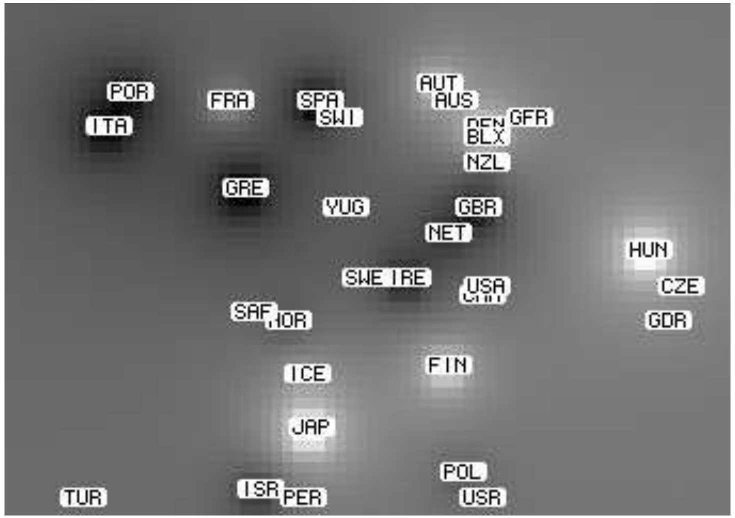


FIGURE 4y Clustered as FIG:S 4.1a & 4.1c. Shows suicide rates 1982-86. (No values for GDR, PER, SAF, TUR, USR, and YUG)

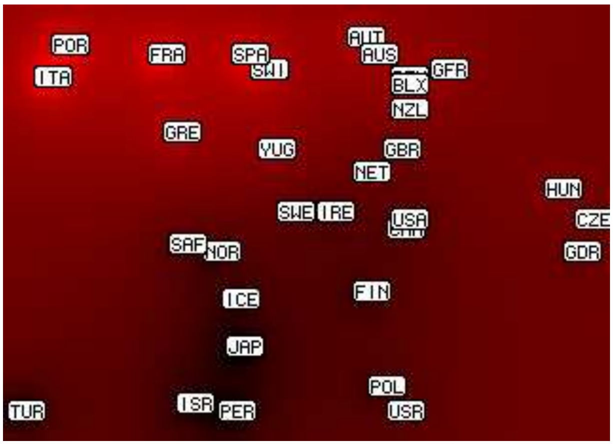

POR
FRA
SPA
AUT
AUS
GFR
ITA
SWI
BLX
NZL
GRE
YUG
GBR
NET
HUN
SWE
IRE
CZE
USA
SAF
NOR
GDR
FIN
ICE
JAP
POL
TUR
ISR
PER
USR

FIGURE 4.2a Monochrome (red) version of FIGURE 4.1a

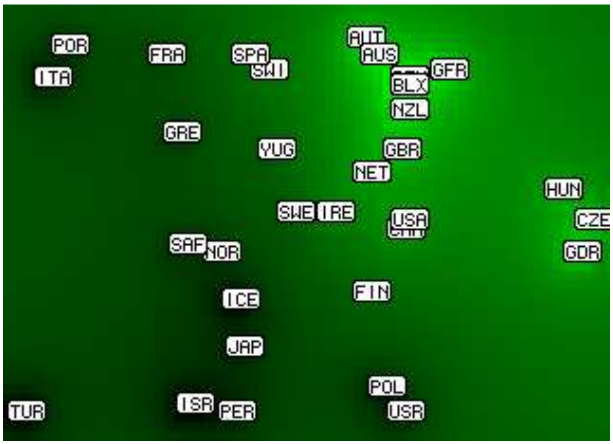

POR
FRA
SPA
AUT
AUS
GFR
ITA
SWI
BLX
NZL
GRE
YUG
GBR
NET
HUN
SWE
IRE
CZE
USA
SAF
NOR
GDR
FIN
ICE
JAP
POL
TUR
ISR
PER
USR

FIGURE 4.2b Monochr. (green) distrib. of beer consumption

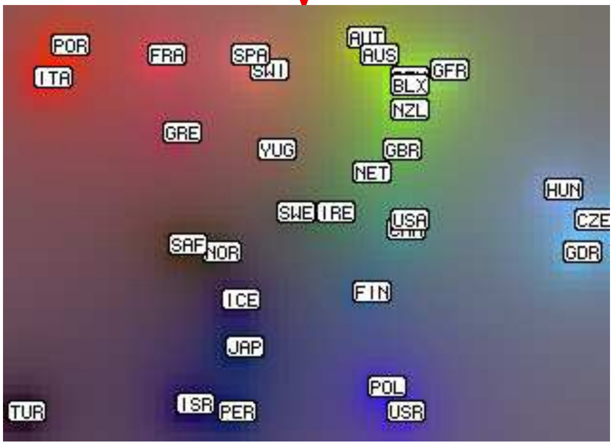

POR
FRA
SPA
AUT
AUS
GFR
ITA
SWI
BLX
NZL
GRE
YUG
GBR
NET
HUN
SWE
IRE
CZE
USA
SAF
NOR
GDR
FIN
ICE
JAP
POL
TUR
ISR
PER
USR

FIGURE 4.2Σ Polychrome synthesis of FIG:S 4.2a, 4.2b, and 4.2c

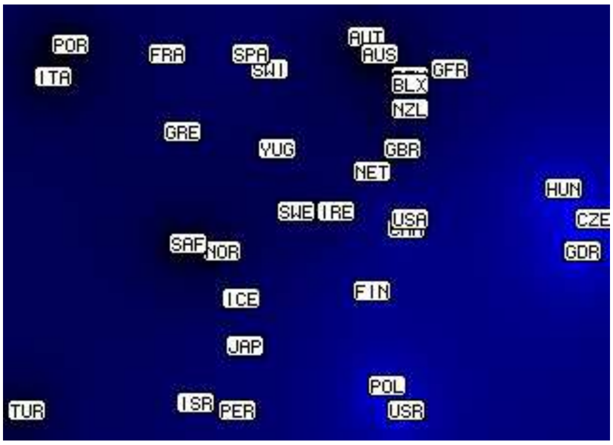

POR
FRA
SPA
AUT
AUS
GFR
ITA
SWI
BLX
NZL
GRE
YUG
GBR
NET
HUN
SWE
IRE
CZE
USA
SAF
NOR
GDR
FIN
ICE
JAP
POL
TUR
ISR
PER
USR

FIGURE 4.2c Monochrome (blue) version of FIGURE 4.1c

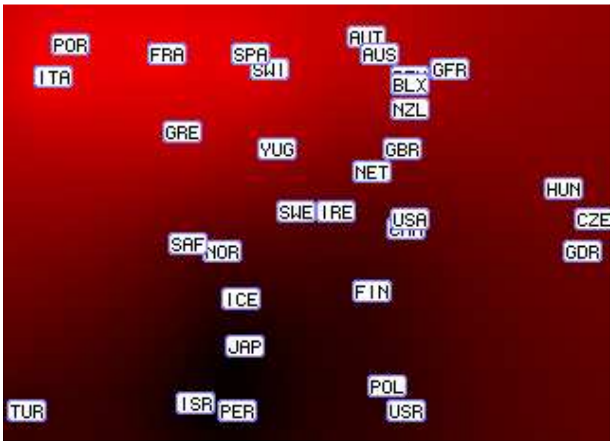

POR
FRA
SPA
AUT
AUS
GFR
ITA
SWI
BLX
NZL
GRE
YUG
GBR
NET
HUN
SWE
IRE
CZE
USA
SAF
NOR
GDR
FIN
ICE
JAP
POL
TUR
ISR
PER
USR

FIGURE 4.3a Smoothed version of FIGURE 4.2a

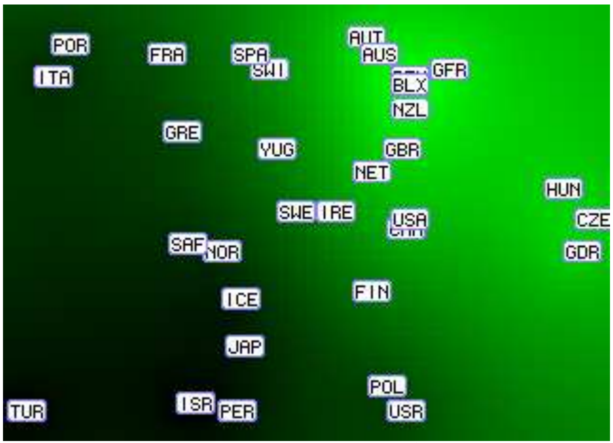

POR
FRA
SPA
AUT
AUS
GFR
ITA
SWI
BLX
NZL
GRE
YUG
GBR
NET
HUN
SWE
IRE
CZE
USA
SAF
NOR
GDR
FIN
ICE
JAP
POL
TUR
ISR
PER
USR

FIGURE 4.3b Smoothed version of FIGURE 4.3b

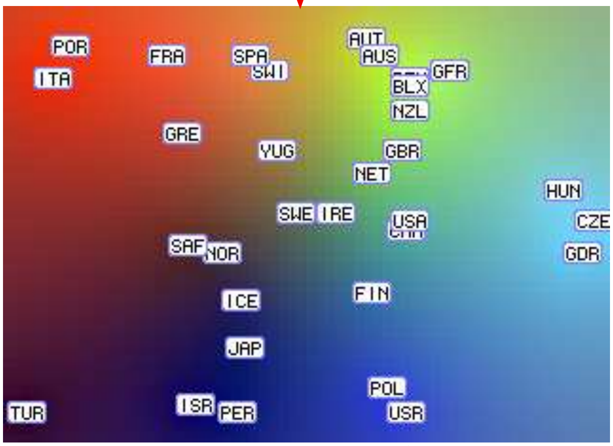

POR
FRA
SPA
AUT
AUS
GFR
ITA
SWI
BLX
NZL
GRE
YUG
GBR
NET
HUN
SWE
IRE
CZE
USA
SAF
NOR
GDR
FIN
ICE
JAP
POL
TUR
ISR
PER
USR

FIGURE 4.3Σ Smoothed version of FIGURE 4.2Σ

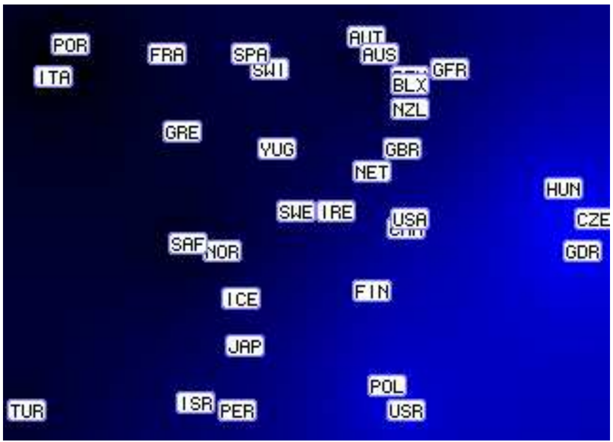
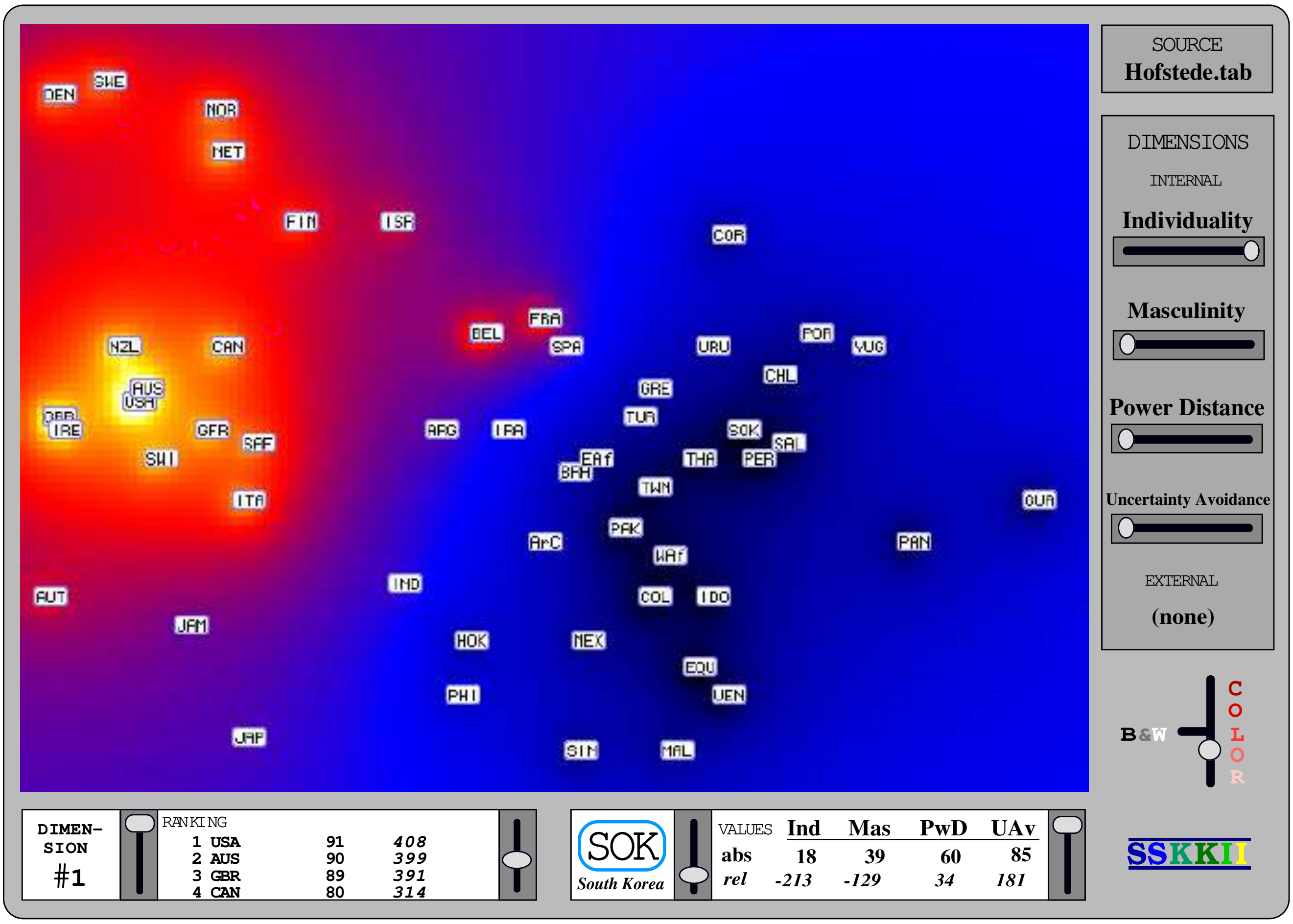

POR
FRA
SPA
AUT
AUS
GFR
ITA
SWI
BLX
NZL
GRE
YUG
GBR
NET
HUN
SWE
IRE
CZE
USA
SAF
NOR
GDR
FIN
ICE
JAP
POL
TUR
ISR
PER
USR

FIGURE 4.3c Smoothed version of FIGURE 4.2c


DEN
SWE
NOR
NET
FIN
ISR
COR
FRA
BEL
SPA
POR
YUG
NZL
CAN
URU
CHL
GRE
AUS
USA
GBR
IRE
GER
SAF
ARG
IRA
TUR
SOK
SAL
EAf
BRA
THA
PER
SWI
TWN
ITA
GUA
PAK
ArC
PAN
WAf
AUT
IND
COL
IDO
JAM
HOK
MEX
EQU
PHI
VEN
JAP
SIN
MAL
B&W
COLOR
SSKKII

SOURCE
Hofstede.tab

DIMENSIONS
INTERNAL
Individuality
Masculinity
Power Distance
Uncertainty Avoidance
EXTERNAL
(none)

DIMENSION #1

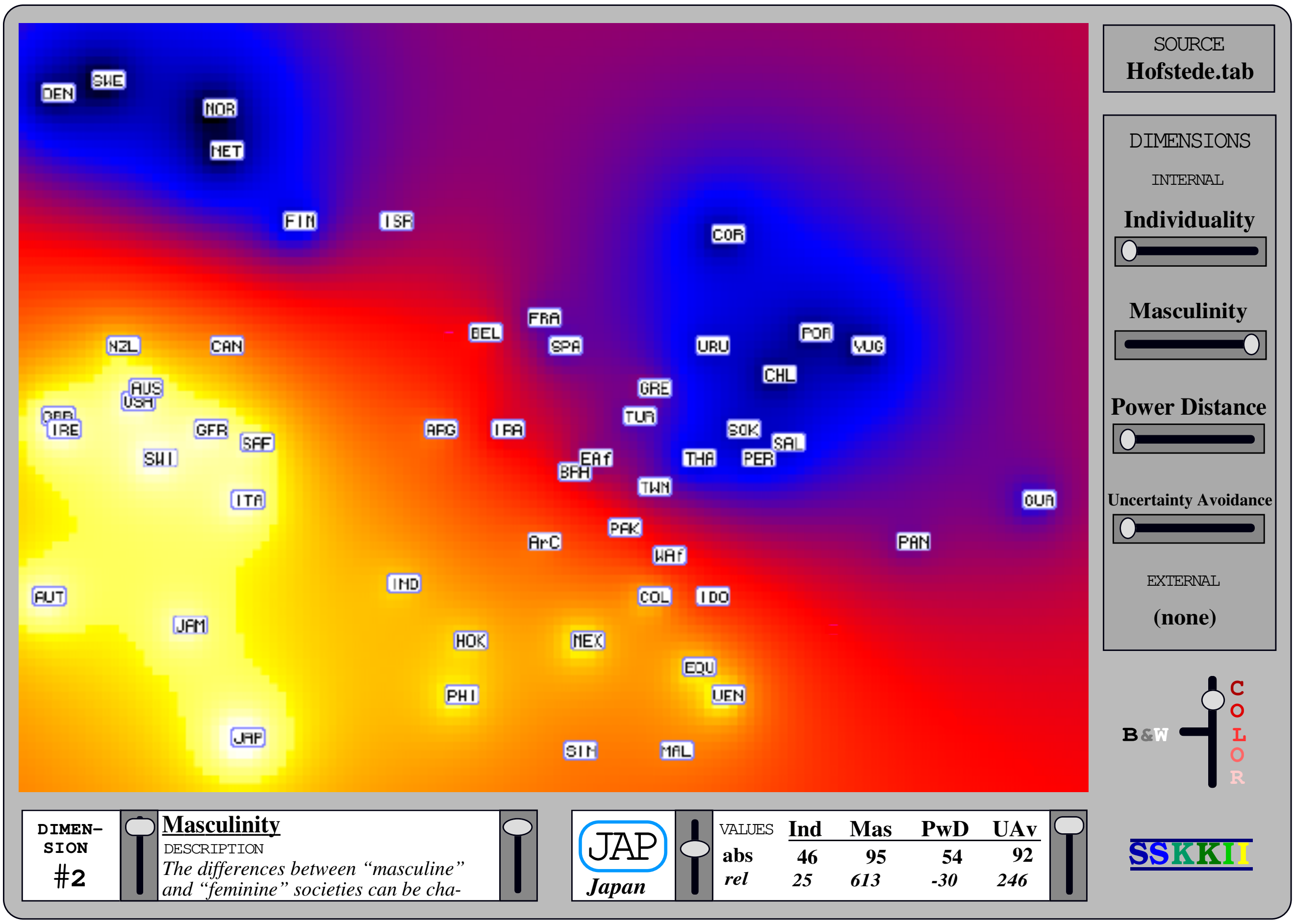
RANKING

1	USA	91	*408*
2	AUS	90	*399*
3	GBR	89	*391*
4	CAN	80	*314*

SOK
South Korea

VALUES	Ind	Mas	PwD	UAv
abs	18	39	60	85
rel	*-213*	*-129*	*34*	*181*

FIGURE 5.1 Example of interface of a tool for information visualization and exploration. The cluster configuration is the result of a CAV analysis of intercultural data from [Hofstede 91]. The current setting shows the value distribution of the feature 'individuality' with low colour smoothing.


DEN
SWE
NOR
NET
FIN
ISR
COR
FRA
BEL
SPA
POR
YUG
NZL
CAN
URU
CHL
GRE
AUS
USA
GBR
IRE
GER
SAF
ARG
IRA
TUR
SOK
SAL
EAf
BRA
THA
PER
SWI
TWN
ITA
GUA
PAK
ArC
PAN
WAf
AUT
IND
COL
IDO
JAM
HOK
MEX
EQU
PHI
VEN
JAP
SIN
MAL
B&W
COLOR
SSKKII

SOURCE
Hofstede.tab

DIMENSIONS
INTERNAL
Individuality
Masculinity
Power Distance
Uncertainty Avoidance
EXTERNAL
(none)

DIMENSION #2

Masculinity
DESCRIPTION
The differences between "masculine" and "feminine" societies can be cha-

JAP
Japan

VALUES	Ind	Mas	PwD	UAv
abs	46	95	54	92
rel	*25*	*613*	*-30*	*246*

FIGURE 5.2 The same interface as in FIGURE 5.1 when indicating the value distribution of the feature 'masculinity' with high colour smoothing.

References

Ahlberg, C. and B. Shneiderman (1994)
"Visual Information Seeking: Tight Coupling of Dynamic Query Filters with Starfield Displays", In *Proceedings ACM CHI'94: Human Factors in Computational Systems.*

Ahlberg, C. and S. Truvé, 1994. "Exploring Terra Incognita in the Design Space of Query Devices", submitted to UIST'94

Allwood, J. and Hagman, J. (1994) "Some Simple Automatic Measures of Spoken Interaction", in *Proceedings from the joint 14th Scandinavian Conference of Linguistics & 8th Conference of Nordic and General Linguistics*, University of Göteborg

Ammerman, A. J. and L. L. Cavalli-Sforza (1984)
The Neolithic Transition and the Genetics of Populations in Europe Princeton Univ. Pr.

Bell, D. A., F. J. McEarlean, P. M. Stewart, and S. J. McClean (1990)
"Application of Simulated Annealing to Clustering Tuples in Databases", in *Journal of the American Society for Information Science*, 41/2 (pp. 98-110)

Brodlie, K.W, L. A Carpenter, R. A. Earnshaw, J. R. Gallop, R. J. Hubbold, A. M. Mumford, C. D. Osland, P. Quarendon (1992)
Scientific Visualization - Techniques and Applications, Springer-Verlag

Cavalli-Sforza, L. L., P. Menozzi, and A. Piazza (1994)
The History and Geography of Human Genes, Princeton, N. J. Princeton Un cop.

Chalmers, M. and P. Chitson (1992) "Bead: Explorations in Information Visualization" in *SIGIR '92, Proceedings from the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* ed. by N. Belkin, P. Ingwersen, and A. Mark Pejtersen IEEE Computer Society Press (pp. 330-337)

Crouch, C. J., and B. Yang (1992)
"Experiments in Automatic Statistical Thesaurus Construction", in *SIGIR '92, Proceedings from the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* ed. by N. Belkin, P. Ingwersen, and A. Mark Pejtersen IEEE Computer Society Press

Eikmeyer, H.-J. (1992)
"Structuring Phonological Space with Self-Organizing Maps", forthcoming in *Papers from the 13th Scandinavian Conference of Linguistics*, Roskilde (ed.) L. Heltoft

Hagman, J. (1994)
Information Visualization and Exploration — Potential and Requirements, SSKKII Technical Report #3, University of Göteborg / Chalmers University of Technology

Hofstede, G. (1991) *Cultures and Organizations*, McGraw-Hill Book Company

Jain, A. K. and Dubes, R. C. (1988)
Algorithms for Clustering Data, Prentice Hall Advanced Reference Series

Kirkpatrick, S., C. D. Gelatt Jr, and M. P. Vecchi (1983)
"Optimization by Simulated Annealing", *Science*, vol. 220 (pp. 671-680)

Lin, X. (1992)
"Visualization for the Document Space", in *Visualization '92*, Conference Proceedings, IEEE Computer Society Press, (pp. 274-281)

Morris, J. and G. Hirst (1991)
"Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text" in *Computational Linguistics*, March '91, (pp. 21-48)

Ritter, H. and T. Kohonen (1989)
"Self-Organizing Semantic Maps", in *Biological Cybernetics*, 61 (pp.241-254)

Salton, G., J. Allan, and C. Buckley
"Automatic Structuring and Rretrieval of Large Text Files", in *Communications of the ACM*, February 1994 / Vol. 37, No.2 (pp. 97-108)

Sammon, J. W. (1969)
"A nonlinear mapping for data structure analysis", *IEEE Transactions on Computers* C 18, (pp.401-409)

SCB (1991)
Statistisk årsbok för Sverige, SCBs Förlag, Stockholm

Tufte, E. R. (1983)
The Visual Display of Quantitative Information, Graphics Press, Cheshire, Connecticut