13 Coloring that Reveals High-Dimensional Structures in Data

Samuel Kaski, Jarkko Venna, and Teuvo Kohonen

When illustrating statistical tables, it is commonplace to visualize different groups of data items with different colors. For example, the World Bank visualizes different groups of economies, viz. low-income, middle-income, and high-income economies, by coloring the countries in each of the three groups with different, manually chosen colors on a world map display.
Similar visualizations are being used so pervasively that the question naturally arises, whether it would be possible to construct such a coloring that the relations of the colors would represent the relations of the clusters or, more generally, so that the *perceptual relations in the colors would reflect the relations between the high-dimensional data items.*
In Section 13.2 we present a solution in the special case that is especially useful for exploratory data analysis: coloring of data sets organized on Self-Organizing Maps. First, however, the basic setting is introduced in a simpler form in which the coloring is chosen interactively.

13.1 Simple method for interactive coloring

The starting point of the coloring is a Self-Organizing Map of the data set. The map can be used to visualize cluster structures in the data on a map display as discussed in Section 11. A sample display that describes the structures of welfare and poverty in the countries of the world is shown in Figure 13 of Section 11, and reproduced in Figure 19.
In the interactive coloring system the user decides, based on the clustering display, how many clusters there are in the data, points out the cluster centers, and chooses colors for them. A sample choice of the centers has been shown in Figure 19. An automatic system [1] may be used to make preliminary choices.
After the cluster centers have been colored, the color is "spread" to the neighborhood of each center; while the color spreads its intensity diminishes according to the distance it has passed. The clustering structure (illustrated in shades of gray in Figure 19) is taken into account when computing the distance: in the clustered areas the intensity diminishes more slowly than in the "ravines" between the clusters. Each location on the map receives the color that is a mixture of the colors that have spread to it from the cluster centers, each weighted by the distance of that unit from the corresponding center.
In the resulting map display (Figure 20) the colors have an intuitive interpretation: Each bright (pure) color corresponds to a certain data type, and mixed colors correspond to intermediate forms.

13.2 Automatic coloring

The coloring described in the previous section was more faithful to the relations between the actual data items than a completely manual coloring, but it was still

not perfect.
Ideally, the colors should be such that the relations between the high-dimensional data items would be reflected as closely as possible in the perceptual relations between their colors. The perceptual differences between colors can be approximated by distances in a color space called CIELab. Therefore, any coloring actually corresponds to a certain mapping of the high-dimensional data set into a smaller-dimensional space, the CIELab color space.
The perceptual color space is unfortunately only three-dimensional, and therefore it is impossible in general to construct a mapping that would preserve *all* of the pairwise distances between the data items. Fortunately, all of the distances are not equally important for the quality of the coloring. Longer distances do not usually need to be represented as accurately as shorter ones. In traditional hierarchical clustering, for example, the relation between data items is represented as their relation in a hierarchical tree of clusters. Distances of items in different clusters will then not be represented individually but in terms of the relation between the clusters.
It has turned out in our studies that it is possible to construct a coloring that concentrates almost entirely on representing the local, small distances, but which still becomes globally ordered. The only additional constraint needed is that no data item that is originally farther away may attain a color that is, intuitively speaking, in between the colors of two close-by data items. Such a color would break the global order. These constraints have turned out flexible enough to allow the mapped colors

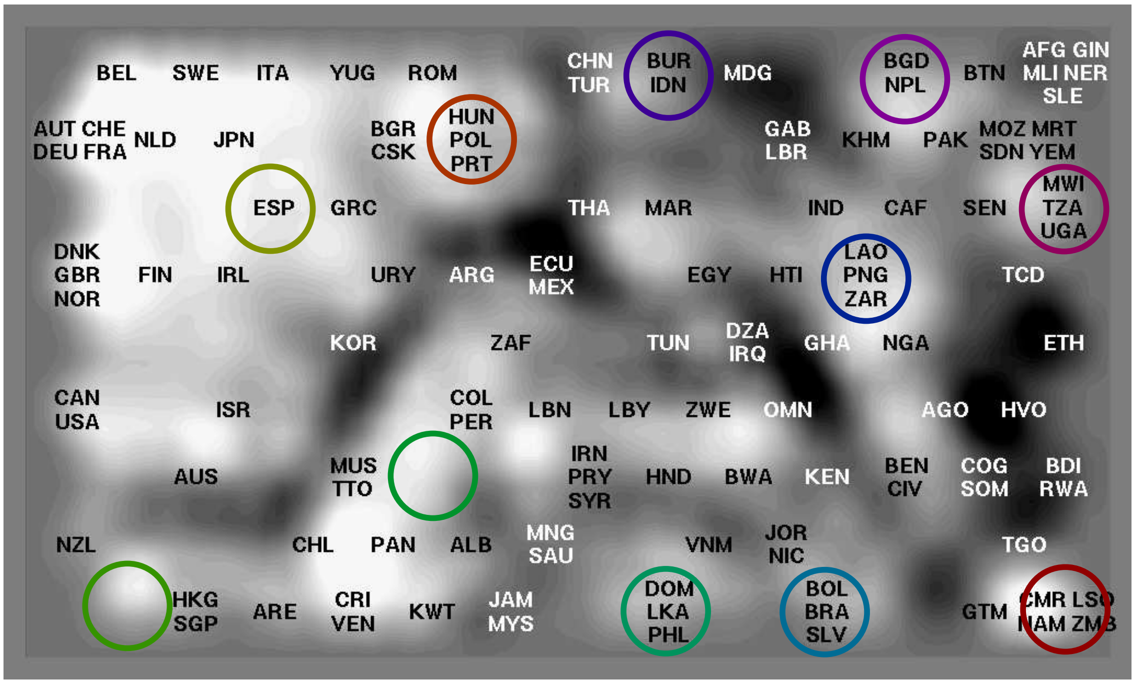

Figure 19: Locations of the chosen cluster centers, shown on top of the clustering diagram formed by the Self-Organizing Map algorithm. The diagram describes different aspects of the welfare and poverty of different countries. Light areas represent areas of a high degree of clustering and dark areas gaps in the degree of clustering. Different types of welfare and poverty are visible as the clustered (light) areas on the map, separated by the dark "ravines".

to fill the available color space reasonably well.
The mapping method has been applied to coloring SOM displays which are well-suited for such coloring. Neighboring map units represent similar data items, and therefore the distances between neighboring units may be regarded as the local ones that will be represented accurately.
The mapping is constructed by requiring that (1) the local distances, i.e. distances between model vectors of neighboring map units, will be preserved as accurately as possible; (2) the model vectors belonging to farther-away map units remain farther away (but their relative order may be arbitrary); and (3) the colors remain within the region of the color space which is representable in the chosen media, for example by the CRT tube. Each of these three conditions was dressed into a term in a cost function, whereafter the minimum of the cost function can be sought with any standard optimization algorithm. Sor far we have used stochastic gradient descent which aids in avoiding local minima.
The result of mapping the model vectors of the SOM of Figure 19 into the CIELab color space is shown in Figure 21. When the map units were colored according to the projection (Figure 22), the differences in the hue of neighboring map units corresponded well with the distances in the original data space, depicted as shades of gray in Figure 19. In addition, the hues became ordered globally; different clustered areas attained different, relatively uniform hues.
The resulting coloring is almost tailored for the human color vision system which

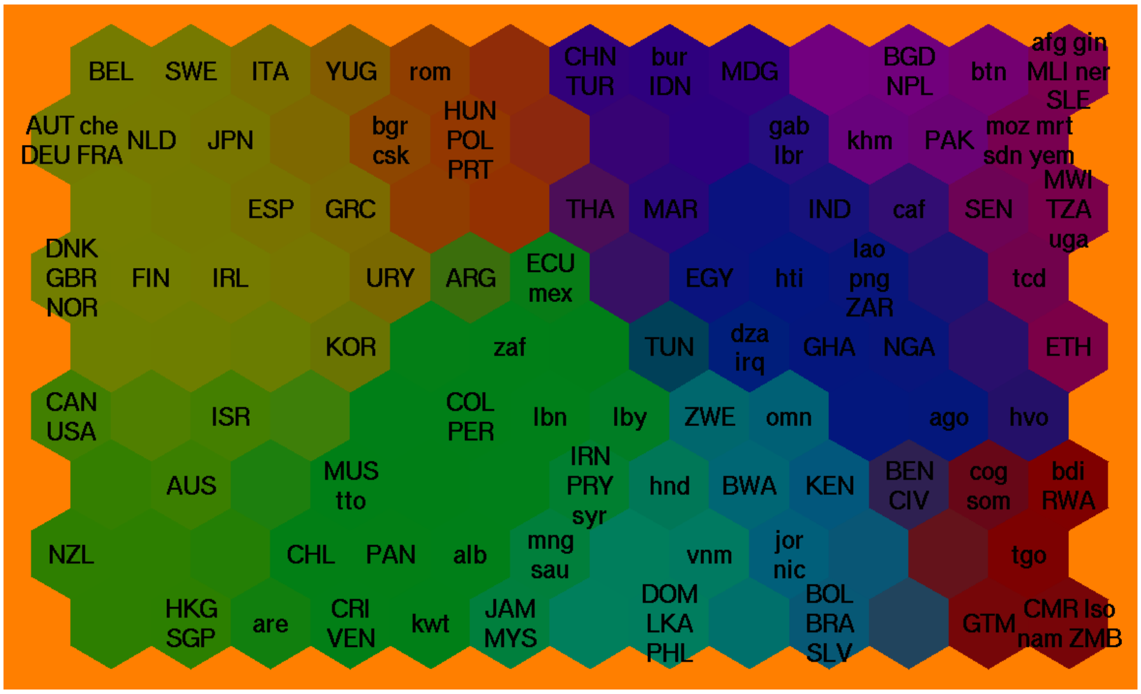

Figure 20: A Self-Organizing Map display in which different, manually chosen clusters have been colored with different colors. The colors are brightest in the cluster centers, and change gradually with increasing distance from the center. The colors change in proportion to the clustering structure so that tight clusters have a relatively homogeneous coloring, and the color changes the more sharply the more clear-cut the border between the clusters is.

is very accurate in detecting differences between the colors of neighboring areas, in this case the neighboring map units.

Relation to possible alternative methods. The traditional multidimensional scaling methods like the Sammon's mapping [2] could in principle be used for projecting the data items into the color space. They do not, however, produce as flexible mappings as our method, since they try to represent *all* of the pairwise distances. The local differences will then necessarily be represented less accurately, except in special cases. Moreover, it may be more difficult to utilize all of the available color space when the mapping is more stiff.

In principle our method could be used to map the data set directly, instead of mapping the model vectors of a SOM computed from the data set. It would, however, be more difficult to define which distances are local enough so that they should be represented accurately. If it is necessary to obtain a characteristic color for each data item then local linear approximations, for instance, may be used to complement the mapping.

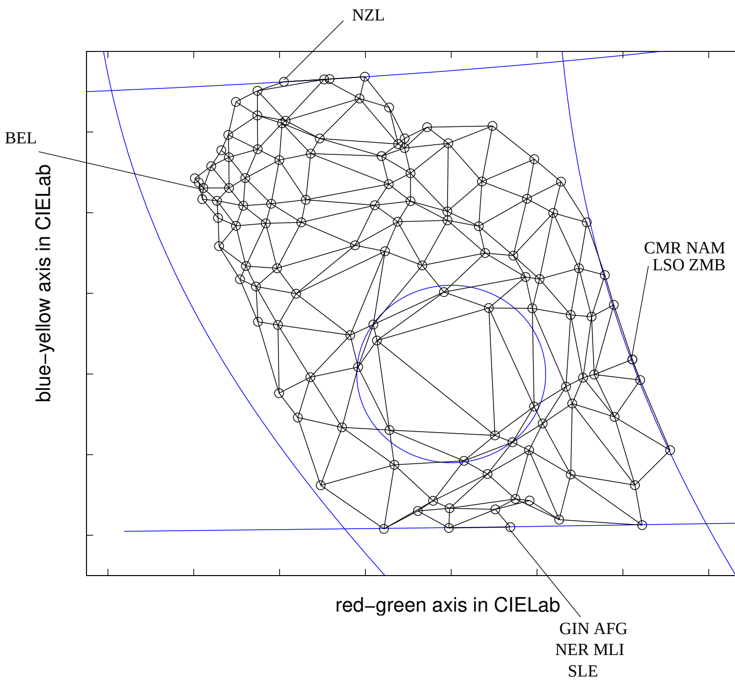

Figure 21: The projection of the model vectors of the SOM of Figure 19 into the CIELab color space. Only part of the space was used to ensure that the obtained colors would differ only in one perceptual quality, the hue. The lightness was fixed to a constant value, which reduces the space into a two-dimensional slice of the original three-dimensional color space, and projections onto non-saturated colors in the middle, encircled in the figure, were discouraged. The small circles denote the projections of the model vectors. Projections of model vectors of neighboring map units have been connected with lines. The long lines delimit the region representable by a typical CRT tube.

13.3 Example: coloring of the world map according to poverty types

The coloring can be even more useful if the original data set can be visualized also in some other manner. Then each data item can be colored with the color that the item has on the SOM display. The welfare and poverty structures can be visualized in a straightforward manner: the countries can be colored according to their welfare or poverty type on a geographic map display (Figure 23).

The result is a display where countries having a similar welfare or poverty type have been colored similarly irrespective of their geographical location. Japan and Australia, for example, are fairly similar to the European countries and the USA and Canada. Countries which belong to very different types than their neighbors pop out strongly, like Japan, Sri Lanka, and Albania.

Visualizations like the one shown in Figure 23 can be very useful if the data has a "natural" ordering like the geographical order here. In fact, *any* order of the data items can be used. For example, if the countries were ordered simply according to the GNP per capita in a statistical table and colored using a SOM, then countries in which the welfare or poverty type is different from the other countries having a similar value of GNP per capita would be clearly discernible based on sharp discontinuities in the coloring.

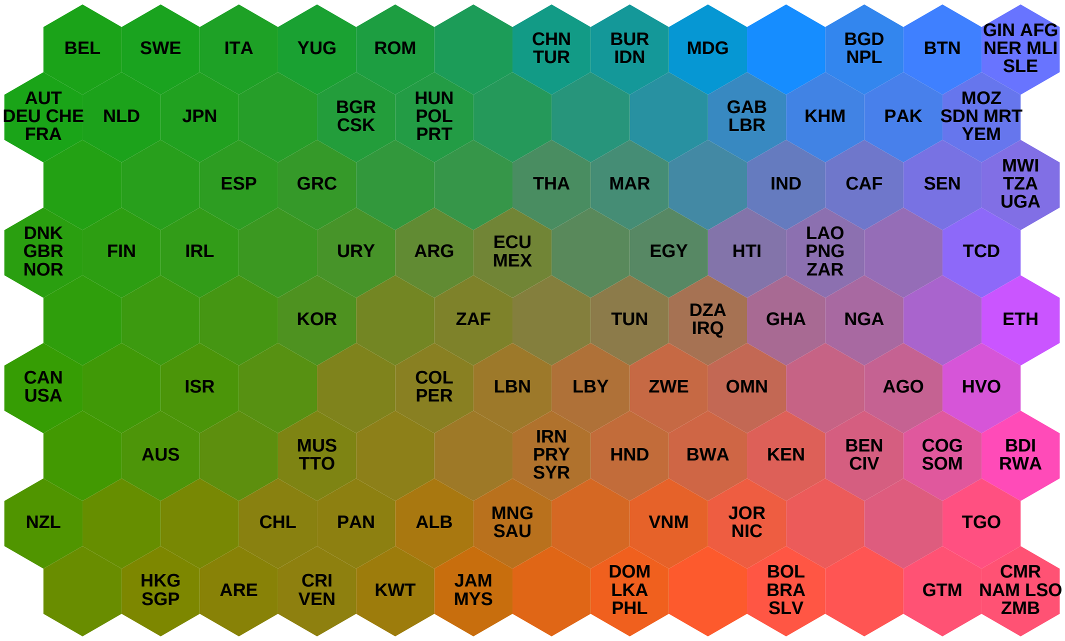

Figure 22: Coloring of the SOM according to the projection shown in Figure 21. The relative differences in the colors of the neighboring map units reflect closely the clustering display in Figure 19. The cluster areas have relatively uniform coloring, and the differences are the larger the steeper the "ravine" between the clusters is.

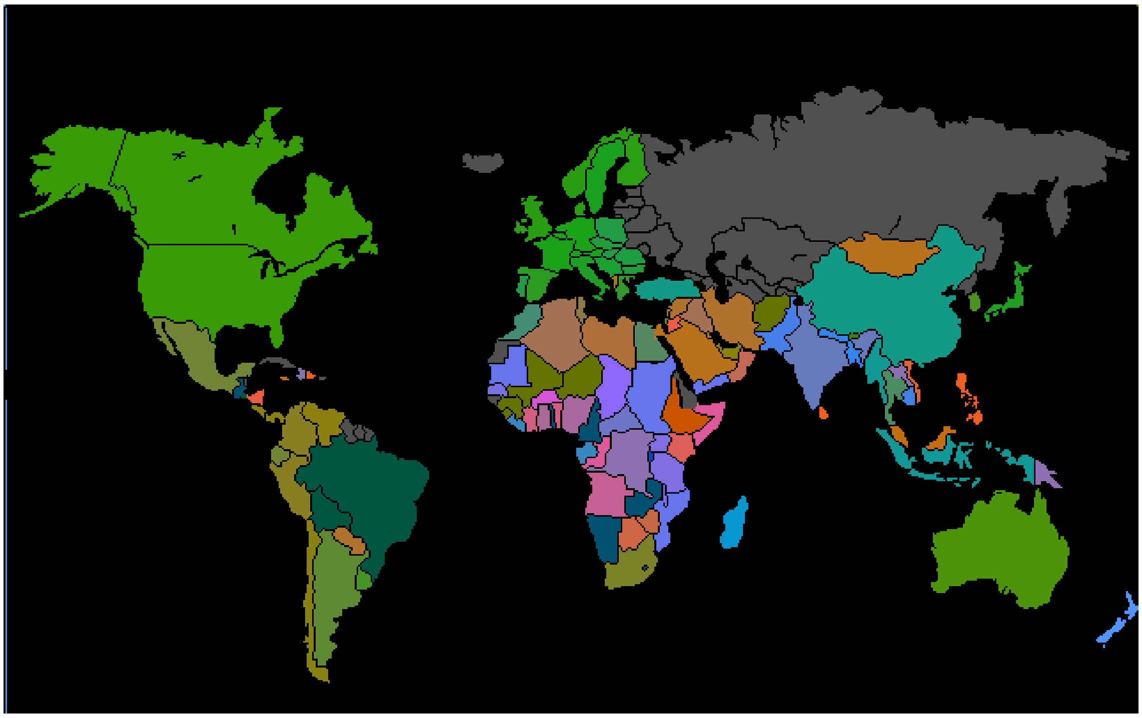

Figure 23: The types of welfare and poverty that the Self-Organizing Map has revealed can be visualized on a geographical world map. Each country is colored according to its color on the SOM display (Figure 22). Countries for which no data was available (like Russia) have been colored with dark gray.

13.4 Conclusions

We have constructed an automatic method for coloring data so that the perceptual properties of the coloring reflect closely the properties of the high-dimensional statistical data. The easily interpretable coloring makes it possible to visualize complex statistical structures automatically for non-experts in statistics.

References

[1] S. Kaski, J. Venna, and T. Kohonen. Tips for processing and color-coding of self-organizing maps. In Guido Deboeck and Teuvo Kohonen, editors, *Visual Explorations in Finance with Self-Organizing Maps*, pages 195–202. Springer, London, 1998.

[2] John W. Sammon, Jr. A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers*, C-18:401–409, 1969.