

# SOM visualization applied for data of countries from four continents

Nguyen Luong Dong [dongnl@pmail.ntu.edu.sg](mailto:dongnl@pmail.ntu.edu.sg)

## Abstract:

Self-organizing map (SOM) is a data visualization technique invented by Professor Teuvo Kohonen which reduces the dimensions of data through the use of self-organizing neural networks. The way SOM goes about reducing dimensions is by producing a map of usually 1 or 2 dimensions which plot the similarities of the data by grouping similar data items together. So SOM accomplishes two things, they reduces dimensions and displays similarities.

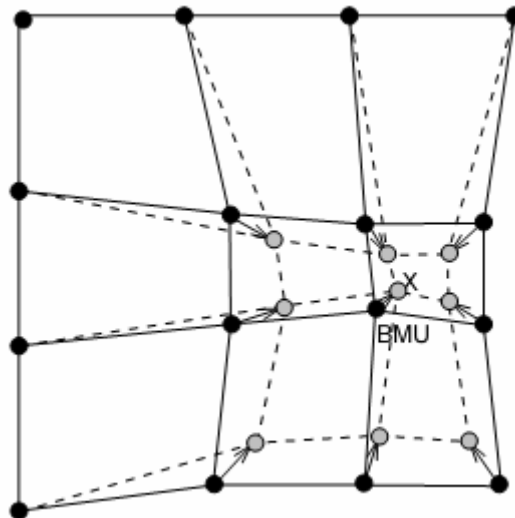
In this paper, we see how similar and different countries of four continent (Europe, America, Asia, and Africa) look like. The data to be visualized includes four categories: People, Economy, Communication, Transportation. There are categories in which each continent is very different from others while in some categories continents are quite similar. All raw data is collected from CIA Factbook 2006.

There are many software that implement SOM, but for reason of simplicity here we use SOM toolbox developed for MATLAB by Helsinki University of Technology.

## 1. Introduction to SOM

A SOM contains neurons which are organized on a regular low-dimensional grid. Each neuron is represented by a  $d$ -dimensional weight vector  $m=[m_1, \dots, m_d]$ . The neurons are connected to adjacent neurons by a neighborhood relation describing structure of the map.

The SOM training algorithm resembles vector quantization algorithm, such as  $k$ -means. The significant distinction is that in addition to the best matching weight vector, also its topological neighbors on the map are updated: the region around the best matching vector is stretched towards the presented training sample:



The end result is that the neurons on the grid become ordered: neighboring neurons have similar weight vectors.

### Sequential training algorithm:

The SOM is trained iteratively. In each training step, one sample vector  $x$  from the input data set is chosen and the distance between it and all the weight vectors of the SOM are calculated using some distance function. The neuron whose weight vector is closest to the input vector  $x$  is called the Best Matching Unit (BMU), denoted here by  $c$ :

$$\|x - m_c\| = \min \{ \|x - m_i\| \},$$

Where  $\| \cdot \|$  is the distance measure (maybe Euclidian distance).

After finding the BMU, the weight vectors of the SOM are updated so that the BMU is moved closer to the input vector in the input space. The topological neighbors of the BMU are treated similarly towards the input vector.

The SOM update rule for weight vector of unit  $I$  is

$$M_i(t+1) = m_i(t) + \alpha(t)h_{ci}(i)[x(t) - m_i(t)],$$

Where  $t$  denotes time. The  $x(t)$  is an input vector randomly drawn from the input data set at time  $t$ ,  $h_{ci}(t)$  is the neighborhood kernel around the winner unit  $c$  and  $\alpha(t)$  is the learning rate at time  $t$ :

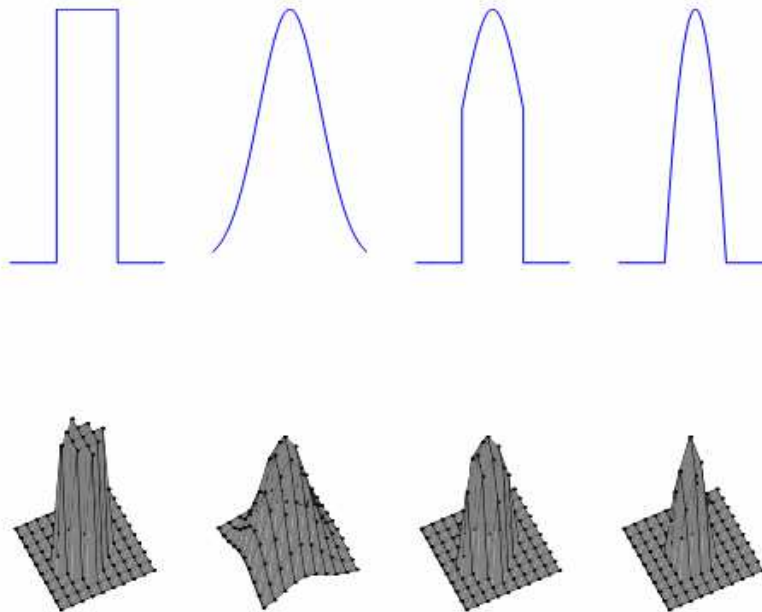


Figure 4: Different neighborhood functions. From the left 'bubble'  $h_{ci}(t) = 1(\sigma_t - d_{ci})$ , 'gaussian'  $h_{ci}(t) = e^{-d_{ci}^2/2\sigma_t^2}$ , 'cutgauss'  $h_{ci}(t) = e^{-d_{ci}^2/2\sigma_t^2} 1(\sigma_t - d_{ci})$ , and 'ep'  $h_{ci}(t) = \max\{0, 1 - (\sigma_t - d_{ci})^2\}$ , where  $\sigma_t$  is the neighborhood radius at time  $t$ ,  $d_{ci} = \|\mathbf{r}_c - \mathbf{r}_i\|$  is the distance between map units  $c$  and  $i$  on the map grid and  $1(x) = 0$  if  $x < 0$  and  $1(x) = 1$  if  $x \geq 0$ . The top row shows the function in 1- and the bottom row on a 2-dimensional map grid. The neighborhood radius used is  $\sigma_t = 2$ .

The neighborhood kernel is a non-increasing function of time and of the distance of unit  $i$  from the winner unit  $c$ . It defines the region of influence that input sample vector has on the SOM.

The training is usually performed in two phases. In the first phase, relatively large initial learning rate  $\alpha_0$  and the neighborhood radius  $\sigma_0$  are used. In the second phase, both learning rate and neighborhood radius are small right from the beginning. This procedure corresponds to first tuning the SOM approximately to the same space as the input data and then fine-tuning the map.

### **Interpretation:**

There are two ways to interpret a SOM. Because in the training phase weights of the whole neighborhood are moved in the same direction, similar items tend to excite adjacent neurons. Therefore, SOM forms a semantic map where similar samples are mapped close together and dissimilar apart.

The other way to perceive the neuronal weights is to think them as pointers to the input space. They form a discrete approximation of the distribution of training samples. More neurons point to regions with high training sample concentration and fewer where the samples are scarce.

## **2. SOM Toolbox**

SOM Toolbox is a software library for Matlab implementing the Self-Organizing Map (SOM) algorithm.

Matlab has been steadily gaining popularity as the "language of scientific computing". For quite a while, the Matlab Neural Networks Toolbox has included a couple of functions that are related to the SOM. These are, however, primarily for demonstrations of the self-organization process and, as such, not sufficient for practical data analysis applications. The SOM Toolbox is the first such software for Matlab.

Highlights of the SOM Toolbox include the following:

- \* Modular programming style: the Toolbox code utilizes Matlab structures and the functions are constructed in a modular manner, which makes it convenient to tailor the code for each users' specific needs

- \* Component names, masks and normalizations: to facilitate data mining process, the input vector components may be given names, and different kinds of (reversible) preprocessing operations can be defined for them. Also, the components may be masked, or weighted, according to their relative importance

\* Batch or sequential training: in data analysis applications, the speed of training may be considerably improved by using the batch version. There are also other training variants, like supervised SOM.

\* Map dimension: maps may be N-dimensional --- although visualization is not supported when  $N > 2$

\* Advanced graphics: building on the Matlab's strong graphics capabilities, attractive figures can be easily produced

\* GUIs: there are also some graphical user interfaces, although the use of command line versions of the functions is strongly recommended

### 3. Data preparation

**Data source:** CIA World Factbook is an annual publication of the Central Intelligence Agency (CIA) of the United States with almanac-style information about the countries of the world. It is prepared by the CIA for the use of U.S. government officials, and its style, format, coverage and content are primarily designed to meet their requirements. However, it is frequently used as a resource for student papers, web sites and non-governmental publications.

**Data preparation:** The data to be visualized in the paper is prepared from raw data that is collected from the Factbook. The prepared data consists of information in four categories about many countries in four continents. For example, following are some first lines of input data in People category:

```
6
#n Total_fertility_rate    Birth_rate    Death_rate    Infant_mortality_rate
Life_expectancy_at_birth    Population
6.69  46.60  20.34  160.2343.34  31056997    ASA Afghanistan
2.03  15.11  5.22   20.75  77.43  3581655    EUR Albania
1.89  17.14  4.61   29.87  73.26  32930091    AFR Algeria
...
```

in which 6 is the dimension of input vector. The number of countries is the size of sample data, ranging approximately from 150 to 200 samples for each category. The four continents to be considered are Europe (EUR), America (AME), Asia (ASA), and Africa (AFR). In some input, WOR denotes World data, and EU denotes European Union data.

Note that there are some exceptions in the input data. Continent Australia consists of too few countries to be considered. And some too small countries and areas are not labeled.

We prepare data in four categories: People, Economy, Communication, and Transportation with the following dimensions.

**People:** *Total\_fertility\_rate*      *Birth\_rate*      *Death\_rate*  
*Infant\_mortality\_rate*      *Life\_expectancy\_at\_birth*      *Population*

225 countries and areas' data.

**Economy:** *PPP-per-capita Labour-force Unemployment-rate Inflation*  
*Electricity-Production Electricity-Consumption ExportImport*

165 countries and areas' data.

**Communication:** *Telephone-mainline-use Telephone-mobile-cellular*  
*Internet-users Internet-hosts*

215 countries and areas' data.

**Transportation:** *Airport Roadway Railway*

136 countries and areas' data.

## 4. Visualization

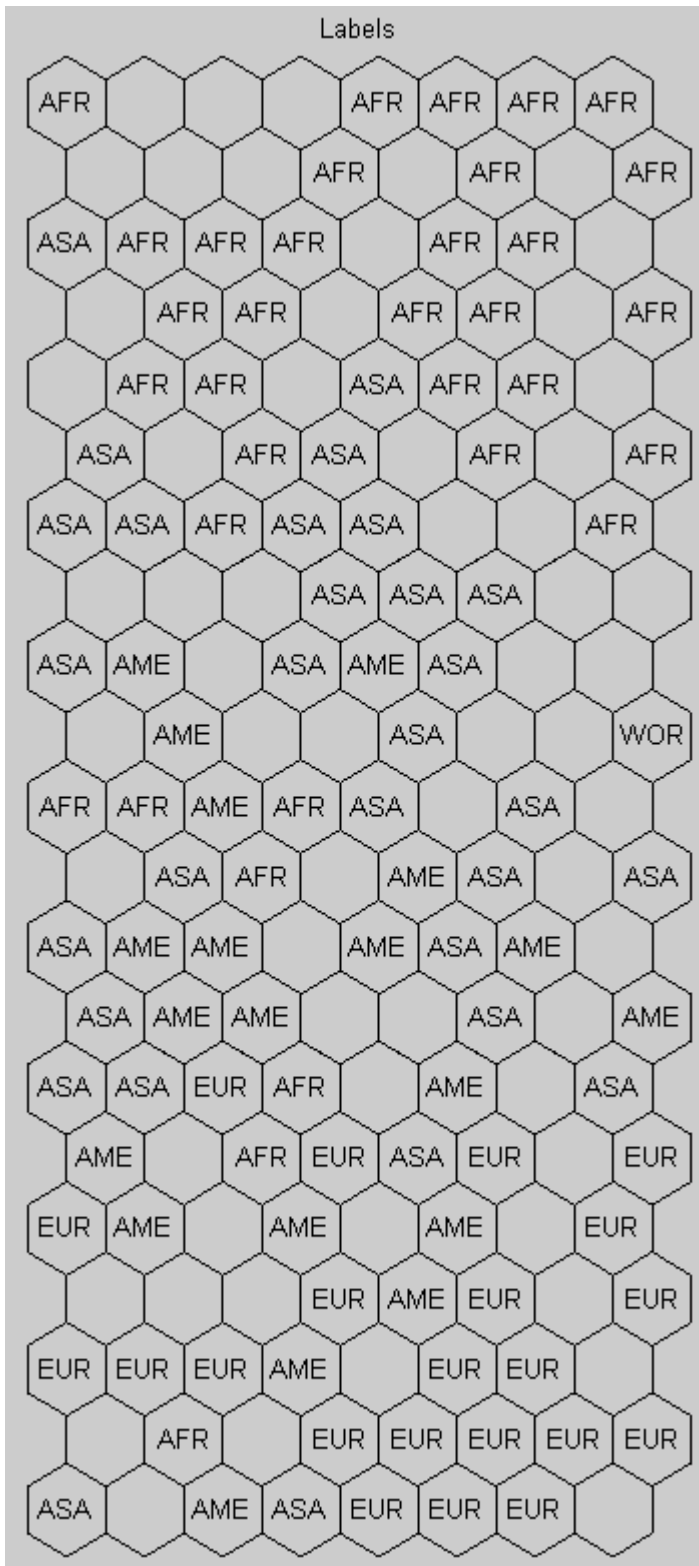
### MATLAB code:

```
% make the data
sD=som_read_data(<filename>);
sD=som_normalize(sD,'var');

% make the SOM
sM=som_make(sD,'munits',<unit_size>,'mapsize','big');
sM=som_autolabel(sM,sD,'vote');

%visualization
som_show(sM,'umat','all','empty','Labels','norm','d');
som_show_add('label',sM,'subplot',2);
```

**a. People category:**



Analyzing SOM in People category:

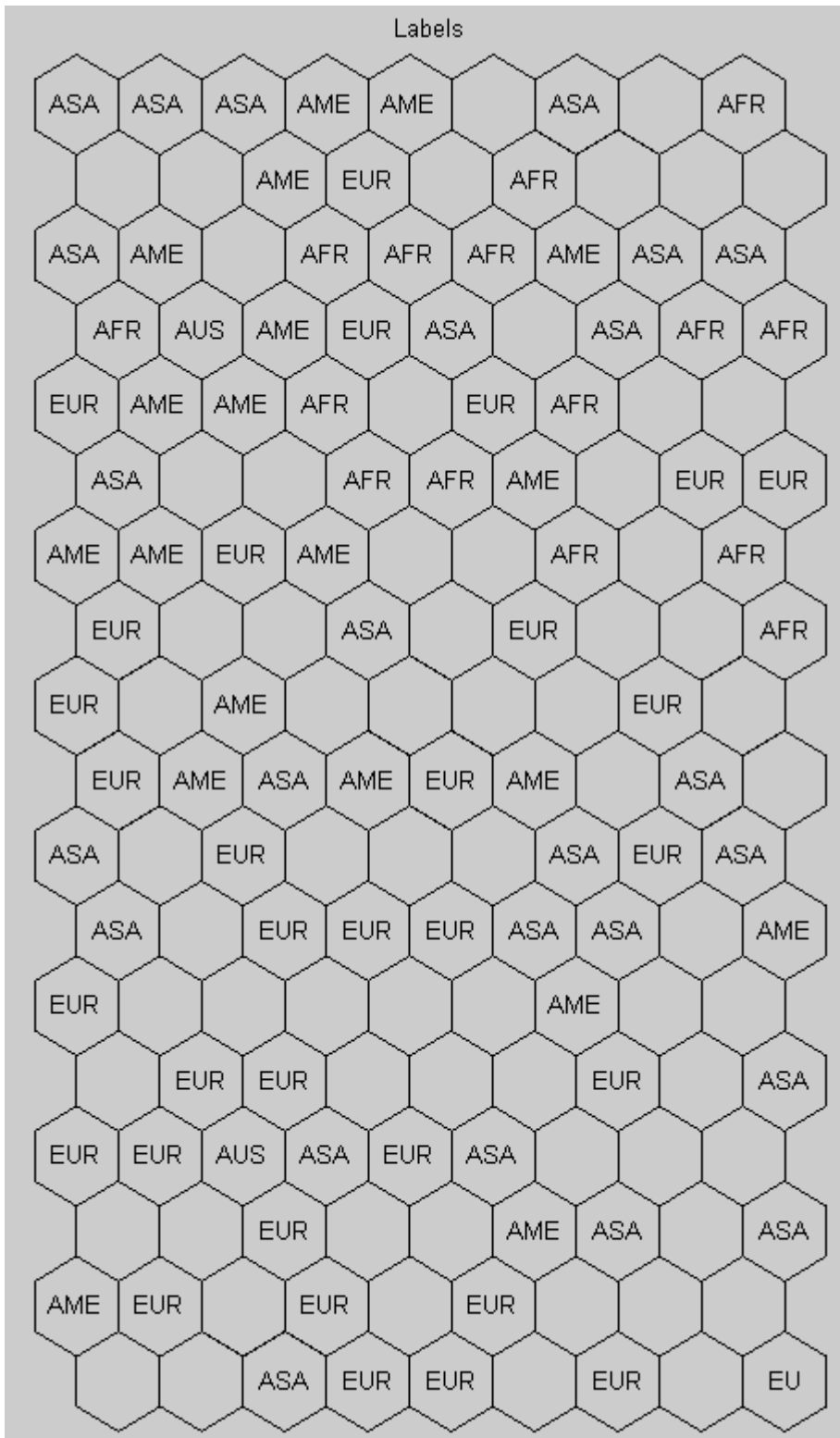
**People:** *Total\_fertility\_rate*      *Birth\_rate*    *Death\_rate*  
          *Infant\_mortality\_rate*      *Life\_expectancy\_at\_birth*    *Population*  
225 countries and areas' data.

The continent has the most similar countries is Europe, with most of its countries lie at bottom right of the map. Africa is quite similar one too. Lying between Europe and Africa is America. It is easy to realize that America is much closer to Europe than Africa is.

At last and rather interesting is Asia. The continent ranges nearly the whole map. This suggests that Asia is the most diversified continent in People category. This understanding might comes from the fact that Asia is the most largest continent with a lot of countries spreading from West to East Asia, from Central to South Asia. These countries diverge very much in geography, in culture, and of course people.

In general, we could draw a conclusion that in the domain of People, Europe and Africa are different with each other, quite similar among their own countries, Asia is distinctly the most diversified, America lies among and is somewhat similar to all three other continents.

**b. Economy category:**



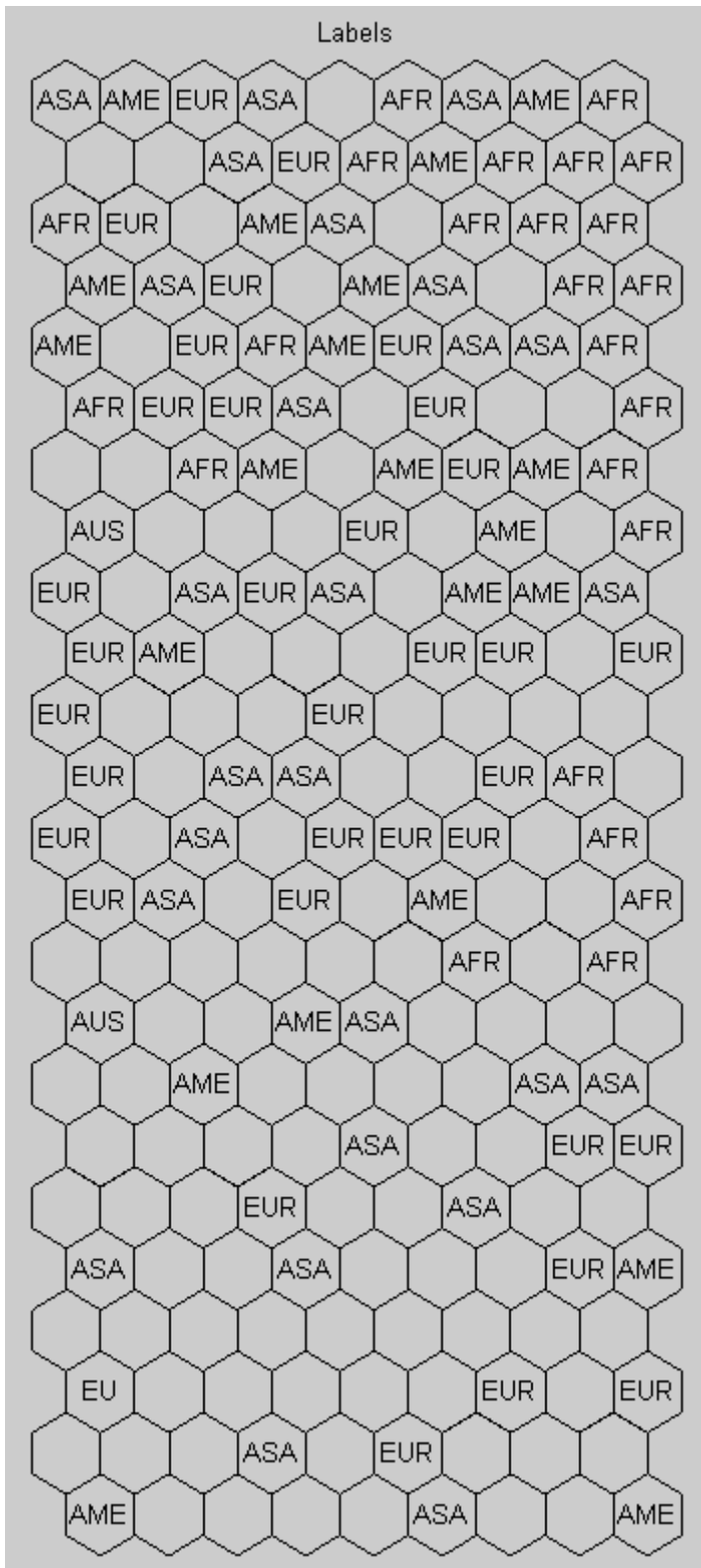
Analyzing Economy category:

**Economy:** *PPP-per-capita Labour-force Unemployment-rate Inflation*  
*Electricity-Production Electricity-Consumption ExportImport*  
165 countries and areas' data.

While People category's SOM shows how continents are different from others, Economy category's map indicates that many countries among other continents may look likely. This suggests that every continent has both rich countries and poor ones.

Of course, there are difference such as most countries which lie on top are from Asia and Africa while most ones below are from Europe and some are from America or Asia. Few countries which are at the bottom of the SOM might be countries from the EU, North America, and East Asia (Japan, South Korea,...)

**c. Communication category:**



Analyzing Communication category:

**Communication:** *Telephone-mainline-use*      *Telephone-mobile-cellular*  
*Internet-users*      *Internet-hosts*

215 countries and areas' data.

Compared to Economy and even People category's SOM, Communication category's one is the most separate. Few developed countries from America, Europe, Asia with good communication condition lie at bottom of the map while a lot of countries from Asia, Africa, America and Europe are at top of the map. When it comes to telephone and internet, surely the United State, the EU and East Asia (Japan, South Korea,...) are the leading countries and areas which we could see at bottom left of the above SOM.



in the case of People category. Nevertheless, most of the times, for instance, in the domain of Economy or Transportation, a lot of countries from different continents interfere in other ones. This fact shows that there is much similarity among continents.

Using SOM as a visualization method shows many interesting facts about different countries from continents in the world. Though SOM method should not be used for the purpose of classification, it sometimes reveals us useful “knowledge” that is not easy to realize in multi-dimension space.