CS-BIGS 3(1): 48-59
© 2009 CS-BIGS

http://www.bentley.edu/csbigs/documents/hua.pdf

Exploratory Analysis of CIA Factbook Data Using Kohonen Self-Organizing Maps

Guangying Hua
Bentley University, USA

Maria Skaletsky
Bentley University, USA

Kimberly Westermann
Bentley University, USA

A visual country comparison is of great importance to research and practice. The method of Kohonen Self-Organizing Maps (SOM) is able to present the data in a visual map and at the same time tries to maintain the topological features of the data. We employ SOM and use data from the 2007 Central Intelligence Agency (CIA) World Factbook to identify what patterns exist between selected countries and their people, economies, communications, and defense forces. Examining 22 indicator variables on 190 countries, we generate a SOM as a means to better understand plausible groupings between countries of the world. The SOM allows us to see a clear visualization of country characteristics by producing multicolored two dimensional hexagonal lattices. In the analysis of these lattices, we identify a tendency for countries to cluster around their particular geographical location – and find that each of these identified clusters has pronounced characteristics associated with each. Extensions of this work over several years may help those interested in world demographics to see the shift in countries over time. This article is accessible to readers with an intermediate level of statistics.

Keywords: *Kohonen maps; Self-Organizing maps; Clustering; CIA World Factbook*

1. Introduction

Country comparison is of great importance to the research and practice, since it not only gives us information about the relative position of a particular country in the world, but also shows the difference among countries. Based on different criteria, the comparison might be different. This paper seeks to use Kohonen Self-Organizing Maps (SOM) in order to identify global patterns of people, economies, communication, and defense for 180 countries in the year 2007 using publicly available Central Intelligence Agency (CIA) data. We expect these findings to inform several audiences. Firstly, this analysis uses publicly available data and provides intimate details of the computation of the maps; therefore this study may be beneficial to those readers interested in learning and/or replicating a study using the Kohonen SOM technique. This analysis may also be used to supplement current studies focused on the environment, globalization, and the incidence of disease or other current global aspects by suggesting common characteristics among specific countries.

Prior research using Kohonen maps to evaluate global data is diverse. Without claiming to exhaustively cite all such research, we mention the following two relevant contributions. Kaski and Kohonen (1995) examine the standard of living in different countries using Kohonen Maps. Deichmann et al. (2007) also use this technique to examine the evolution and determinants of the international digital divide. Nguyen et al. (2008) use a Kohonen map to explore the relative living standards levels of provinces in Vietnam.

2. Methodology

What is a Self Organizing Map?

The competitive SOM methodology is an exploratory data analysis technique, considered to be a special case of a neural network. SOMs are able to project multiple dimension data onto a typically two dimensional hexagonal lattice to enable a clear visualization of the data. This method is suitable for structured statistical data and allows for the identification of groups (in this case, made up of countries) that share similar characteristics. Thus, the Kohonen map, one type of SOM, may also be compared with a cluster analysis, in that SOM serves as an effective mechanism for clustering data. Deichmann et al. comment that "The advantage of the Kohonen approach is the self organizing feature of the map, a very powerful property that makes estimated components vary in a monotonic way across the map." (2007). This distinctive self-organizing property will be discussed further later in the paper.

Kohonen Networks Algorithm

SOM maps are based upon competitive learning, where the output nodes compete among themselves to become the winning node (the node with the best value for a particular score function). In Kohonen learning, the winning node becomes the center of a neighborhood and attracts similar neurons in the center's immediate vicinity. The nodes in the neighborhood will "learn" and adjust their weights using a linear combination of the input vector and the current weight vector to improve the score function. The Kohonen algorithm proceeds as follows:

Consider an input neuron vector $X_n = X_{n1}, X_{n2}, ..., X_{nm}$ and assume that the weights of the neurons, indicated by $\omega_j = \omega_{1j,}\omega_{2j},...,\omega_{mn},$, are initialized typically to small random values. During the competition phase, for each input vector x_i , each output node j and weight vector ω_j, the algorithm calculates the value $D(\omega_j, x_i)$ of the function score (typically using Euclidean Distance) and identifies the map position (the best matching unit BMU) where $D(\omega_j, x_i)$ is optimal (smallest, typically). In the cooperation phase the algorithm identifies all output nodes j within the winning neighborhood for a neighborhood size **R**. Finally, in the adaptation phase, the algorithm adjusts the weights $\omega_{ij,new} = \omega_{ij,current} + \eta(x_i - \omega_{ij,current})$. Here η is a monotonically decreasing learning coefficient. It adjusts the learning rate for a particular neighborhood size as appropriate and stops once the termination criteria are met (the algorithm converges). The algorithm converges when little or no change occurs in the vector of weights. Once the algorithm has converged, the estimated components in the vector of weights will organize themselves on the hexagonal lattice in an ordered fashion (Larose, 2005). Each neuron has actually two positions: one in the input space – the prototype vector – and another in the output space, on the map grid. Thus, SOM is a vector projection method defining a nonlinear projection from the input space to a lower-dimensional output space. On the other hand, during the training the prototype vectors move so that they follow the probability density of the input data. Thus, SOM is also a vector quantization algorithm. SOM is thus an algorithm that combines these two tasks: vector quantization and projection.

How is a Kohonen Map trained using Matlab?

There are several packages to perform the SOM algorithm - online free tools (i.e. Matlab/SOM Toolbox[1]) and commercial packages (i.e. Synapse, http://www.peltarion.com/WebDoc/index.html). The Matlab toolbox is one of the more powerful SOM tools available, producing detailed and clear graphs. The SOM toolbox has many useful functions which enable users to customize their SOM. For example, the SOM toolbox allows the user to train the SOM using different network typologies and parameters and allows the visualization of U-matrices, component planes (to be defined below), cluster color coding and color linking between the SOM

[1] We use the term Matlab and SOM toolbox interchangeably throughout the paper. Matlab is the specific software package and SOM toolbox is a function package for the software.

and other visualization methods. Moreover, the graphical interface of SOM in Matlab, as shown in Appendix 1 is easy to learn and is user friendly (Vesanto, et al., 1999). In this paper, we run a program based on the SOM toolbox, which can be downloaded at no cost from the website http://www.cis.hut.fi/projects/somtoolbox/.

In this paper, the neurons of the map are arranged on a hexagonal lattice, which makes the maps look smoother. The mapping does not considerably suffer even when the number of neurons exceeds the number of input vectors, provided that the neighborhood size is selected properly (Vesanto, 2005).

The number of grid nodes to be used in a SOM-analysis can be considered as a trade-off between representation accuracy and generalization accuracy. A small number of grid nodes will result in a high quantization error and well-defined clusters, while a large number of nodes result in a low quantization error and, in the most extreme case, a cluster for each data sample (Peeters, 2006). We tried several iterations of map sizing to gain a better and more distinguishable U-matrix. A map size with 9 rows and 12 columns was selected as the best map size to represent the data.

Linear initialization is used before training the map. Linear initialization is performed by selecting a mesh of points from the m-dimensional min-max cube of the training data. There are some missing values (less than 10 values per country) in the data and all missing values are handled by simply excluding them from the distance calculation. Batch training is used to train the network. Batch training means that the first part of the SOM algorithm (finding the best matching unit BMU) is applied to the whole training set, and that only after this is the map updated with the net effect applied to all the prototype vectors. Actually, the updating is done by simply replacing the prototype vector with a weighted average over the data points, where the weighting factors are the neighborhood function values.

The batch training algorithm has been proven to be an efficient algorithm used in Matlab and is likely to gain higher quality results compared to normal sequential algorithms (Vesanto, 2005). It is complicated to measure the quality of an SOM. Resolution and topology preservation are generally used to measure SOM quality. There are many ways to measure them. The quantization error (*qe*) and topological error (*te*) are calculated in the Matlab SOM toolbox to measure the quality of the map. The quantization error *qe* is the average distance between each data vector and its BMU, measuring map resolution. The topological error *te* is the proportion of all data vectors for which first and second BMUs are not adjacent units, measuring topology preservation. In the final model, *qe*=2.0492, te =0.0105. Even though according to the literature the quality of a model cannot be evaluated by these two measures only, we still show a good quality for the map. We compare the batch training and sequential algorithms quality measures in table 1. Our model also suggests that the batch algorithm has smaller error values compared to the sequential algorithm.

Table 1. Batch versus Sequential Training Algorithm

Training Algorithm	Quantization error (*qe*)	Topological error (*te*)
Batch	2.0520	0.0105
Sequential	2.4462	0.0211

The Data Set

The statistical indicators used in the analysis were obtained from the CIA Factbook (www.cia.gov). The CIA Factbook contains information on 244 countries/regions of the world and houses approximately 85 descriptive variables for each country. Our final analysis included 190 countries and 22 indicator variables.

Before removing any countries from our data set, we first reduced the number of variables to include in our analysis. Of the 85 descriptive variables available, approximately 50 were easily convertible to numbers for statistical analysis. In addition, current (2007) information may not have been available for all variables. For example we eliminated the variable *illicit drug production* because the information provided per country included a description of past drug cultivation on a per hectare basis for multiple narcotics. In further reducing the number of variables for analysis, we found that not all of the chosen indicators were available for all of the countries. We initially eliminated those variables that had 50 missing values in the country set. For example, we eliminated *oil production* because half of the countries were missing this data point, as a given country may simply not produce oil. The remaining 22 indicators we selected describe various country characteristics of people, economy, communications, and defense forces[2]. Table 2 contains a list of indicator variables including descriptions and descriptive statistics.

[2] Other categories included on the CIA Factbook website and excluded in this analysis include: geography, government, transportation, and transnational issues.

Table 2. Indicator Variables (all transformed with the logarithm function in our analysis)

Variable	Description	Range	Mean	Standard Deviation
Communications				
InternetUser	Number of users within a country that access the Internet *per 100 people*	88.00	20.67	21.92
Tele_Mainline	Total number of main telephone lines in use *per capita*	0.87	0.20	0.20
Tele_Mobile	Total number of mobile cellular telephone subscribers *per capita*	1.72	0.50	0.42
Defense				
MilitaryExpend	Spending on defense programs for the most recent year available as a proportion of GDP	0.11	0.02	0.02
Economy				
Debt	Total *per capita* public and private debt owed to nonresidents repayable in foreign currency, goods, or services. Calculated on an exchange rate basis	338,727	9,827	32,638
Elect_Cons	Total electricity consumed annually plus imports and minus exports, expressed in kilowatt-hours *per capita*	26,999	3,055	4,210
Elect_Prod	Annual electricity generated expressed in kilowatt-hours *per capita*	29.34	3.07	4.67
Exports	Total US dollar amount of merchandise exports on an f.o.b. (free on board) basis *per capita*	63,560	4,349	8,622
GDPAgriculture	Percentage contribution of *agriculture* to total GDP	0.77	0.15	0.15
GDPIndustry	Percentage contribution of *industry* to total GDP.	0.89	0.30	0.15
GDPPPP	A nation's GDP at purchasing power parity (PPP) exchange rates; that is the total value of all goods and services produced in the country valued at prices prevailing in the United States	70,800	12,013	13,443
GDPServices	Percentage contribution of *services* to total GDP	0.90	0.55	0.17
Imports	Total US dollar amount of merchandise imports on a c.i.f. (cost, insurance, and freight) or f.o.b. (free on board) basis *per capita*	53,705	4,402	7,625
Inflation	The annual percent change in consumer prices compared with the previous year's consumer prices	0.53	0.06	0.06
Labor	Total labor force as a percentage of population	59.49	41.38	10.84
Unemployment	Percent of the labor force that is without jobs	0.85	0.13	0.14
People				
Birthrate	The average annual number of births during a year per 1,000 persons in the population at midyear	42.82	22.59	11.45
Popu_Grow	The average annual percent change in the population, resulting from a surplus (or deficit) of births over deaths and the balance of migrants entering and leaving a country	5.00	1.00	1.00
HIVAIDS	An estimate of the percentage of adults (aged 15-49) living with HIV/AIDS. The adult prevalence rate is calculated by dividing the estimated number of adults living with HIV/AIDS at yearend by the total adult population at yearend	10.0	2.00	3.00
LifeExpectancy	The average number of years to be lived by a group of people born in the same year, if mortality at each age remains constant in the future	51.27	66.06	12.78
Fertilitityrate	The average number of children that would be born per woman if all women lived to the end of their childbearing years and bore children according to a given fertility rate at each age	5.93	3.04	1.62
Reserves	The dollar value for the stock of all financial assets that are available to the central monetary authority for use in meeting a country's balance of payments needs as of the end-date of the period specified *per capita*	29,931	1,721	3,523

Of the 244 countries included in the CIA Factbook we excluded those countries with a significant amount of missing variables, very small states and outliers. Some countries were missing few variables, others were missing a substantial number of variables; therefore we excluded from the computation of maps those countries missing more than 10 indicator variables of interest. We excluded very small states, typically "vacation" islands, (i.e. US Virgin Islands, Turks & Caicos) from our analysis.

Prior to analyzing the data, we reviewed the histograms of each indicator variable for normality. We transformed each of our 22 variables with the natural logarithm to ensure a less skewed distribution. We also converted each variable to a per capita basis to take into account the population of each country. This ensures that countries are mapped on a comparable scale. Finally, all variables were standardized (by subtracting the mean and dividing by the standard deviation), ensuring that all variables hold equal importance in the algorithm.

3. Results

Data Visualization (Preliminary analysis)

Our preliminary analysis of the data starts with reviewing the D-matrix and color coding. Distance-matrix (D-matrix) maps indicate the classification mapping of the SOM, which is achieved from training the data. Both the D-matrix and color coding in Figure 1 and Figure 2 show that there are some main clusters in the data based on the size difference and color similarity of nodes. For example in the upper left hand corner of the map in Figure 1 there is a distinct cluster of red, orange and brown and in the lower middle of the map there is a distinct pattern of blue colors.

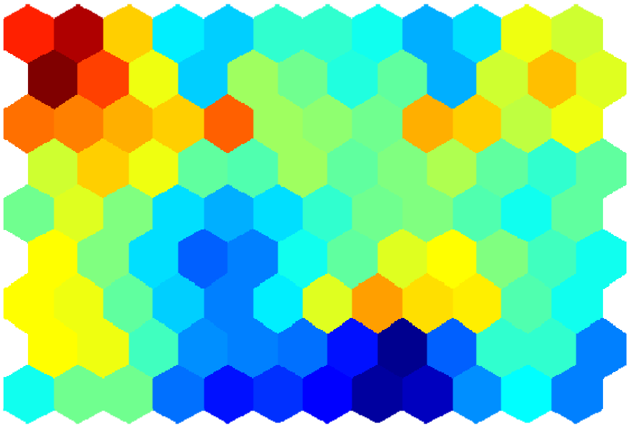


Figure 1. D-matrix with color

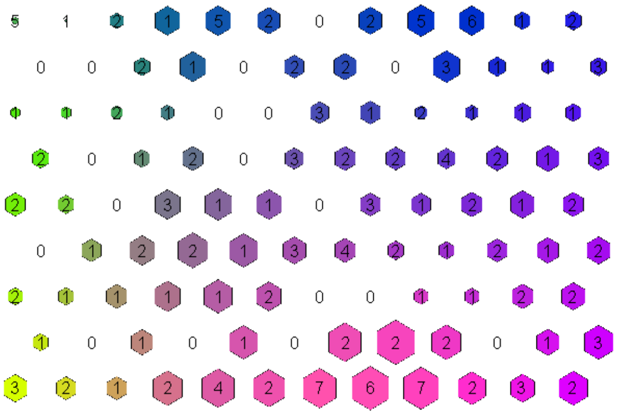


Figure 2. Color coding + distance matrix

Each hexagonal map unit in Figures 1 and 2 represents a SOM node. The number on each cell indicates how many data points it is associated with (that is, how many data points share that cell as Best Matching Unit after training). The size of each node in Figure 2 shows the number of data points associated with the node. The bigger the node size, the more the data points. Very small nodes have very few data points associated with them. Map units with similar prototype vectors have similar colors.

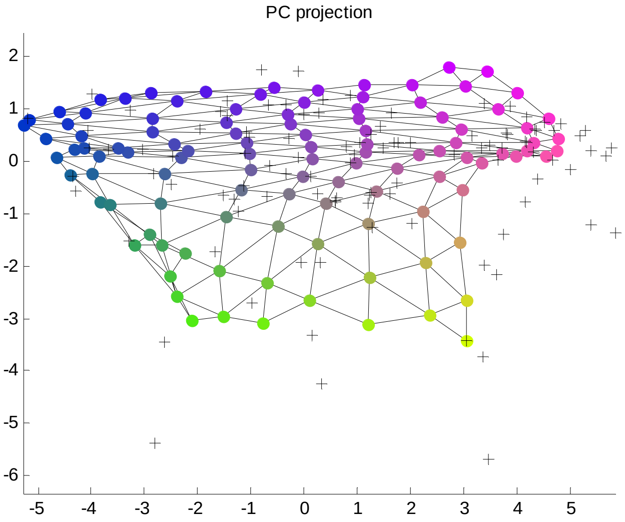


Figure 3. Principal component projection

We note that the SOM methodology is also a projection technique. The principal component (PC) projection is shown in Figure 3. Out of 22 variables, two principal components were extracted. The x and y axes in Figure 3 display the first two principal components and how the data points are projected to both components. Here the colors are used for linking the visualization in Figure 3 to the map plane in Figure 2. The projection figures can be linked to the map planes in Figure 2 using color coding. In Figure 3, the colored nodes represent outputs in the SOM, while the “+” symbols represent the real input data points. The Principal Components projection confirms the existence of clusters in the data.

Initial clustering

A multidimensional data vector is estimated for each cell on the U-matrix, to be further described below. The dimension of these data vectors is equal to the number of variables used in the model. After training, each country is placed on the U-matrix to achieve the smallest distance between its data vector and the estimated prototype vectors of other cells on the map. Multiple countries might be placed in the same cell if their data vectors are closest in distance to the estimated prototype vector of same cell on the map (Deichmann, et al, 2007).

The geometrical relationships between countries on the U-matrix show particular patterns of populations, government, and economic conditions. Countries which are similar on the basis of variables included in the analysis tend to be placed close together on the map, often on the same hexagon. The number of cells of the U matrix is higher than that of the D matrix: intermediate hexagons are inserted in between the initial hexagons of the D matrix. The color of these additional hexagons represents the distance between estimated data vectors for the two neighboring hexagons. For example, a dark red color for an intermediate hexagon indicates that its neighboring hexagons are far apart from each other and a blue color indicates that the distance between these neighboring hexagons is small. The color of a country hexagon represents the average distance between the estimated data vector at that position on the map and the neighboring hexagons (Deichmann, et al, 2007). It indicates how close this country is to countries located in neighboring hexagons - blue indicates that the countries are "close", or more similar, and red means that the countries are "far", or less similar, even if they are located physically close on the map. Therefore, we can determine clustering membership by locating red "walls" between groups of countries.

From the U-matrix (Figure 5) and the D-matrix (Figure 1) we can see that there are at least three clusters in the data. The properties of clusters are different from each other and need to be further analyzed. The K-means non-hierarchical clustering function is used to find an initial partitioning. Cluster validity is important to measure quality of the cluster results. Many cluster validity indices have been proposed. In Figure 3, the Davies-Bouldin clustering index is used to evaluate the quality of k-means clusters. The Davies-Bouldin index is a function of the ratio of the within cluster variation to the between cluster variation (Ingaramo, et al., 2005); small values of the Davies-Bouldin index indicate better clusterings. By using this measure, we want to minimize the within-cluster scatter and maximize the between-cluster separation. The minimization of this index indicates natural partitions of a data set. The smaller the index, the better the partition is (Davies and Bouldin, 1979). The left part of Figure 3 displays the Davies-Bouldin index of each clustering and suggests a solution with 6 clusters. The right part of Figure 3 displays this six-cluster K-means cluster partition.

The two main maps that we will now use further in our analysis are the U matrix (Figure 5) and individual component maps (Figure 6).

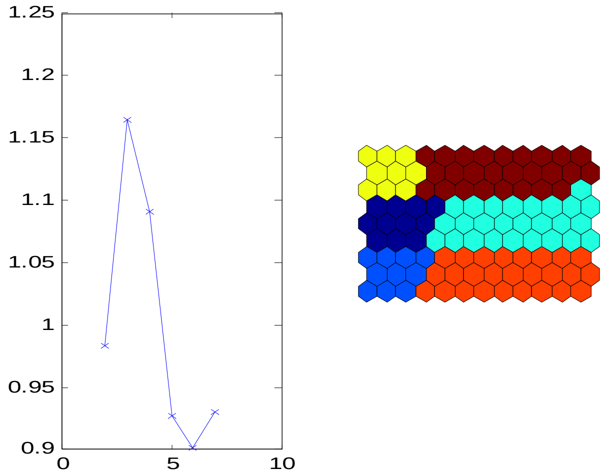


Figure 4. Davies-Bouldin index and K-means clusters

Component maps

Individual component maps (Figure 6) provide clear visualization of the estimated prototype vectors after training. The component maps indicate the estimated values of each individual variable at each map position. Estimated individual indicators are displayed in various shades of reds and blues on the SOM component maps. For example, the HIV rate component map shows mostly a blue color, indicating low estimated values, and a small spot of red color in the top left corner (Figure 7b). This is due to the fact that a few African countries have extremely high estimated HIV infection rates (as high as 25.7% of adult population infected), whereas many countries have HIV infection rates only slightly above 0% as shown in Figure 7(a). Overall, the U-matrix shows measures of proximity between the countries and the interpretation of meaning for moving across the map is derived from the component

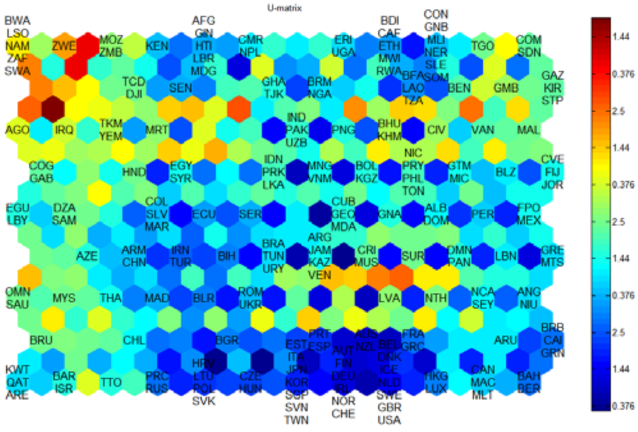


Figure 5. Kohonen U-matrix

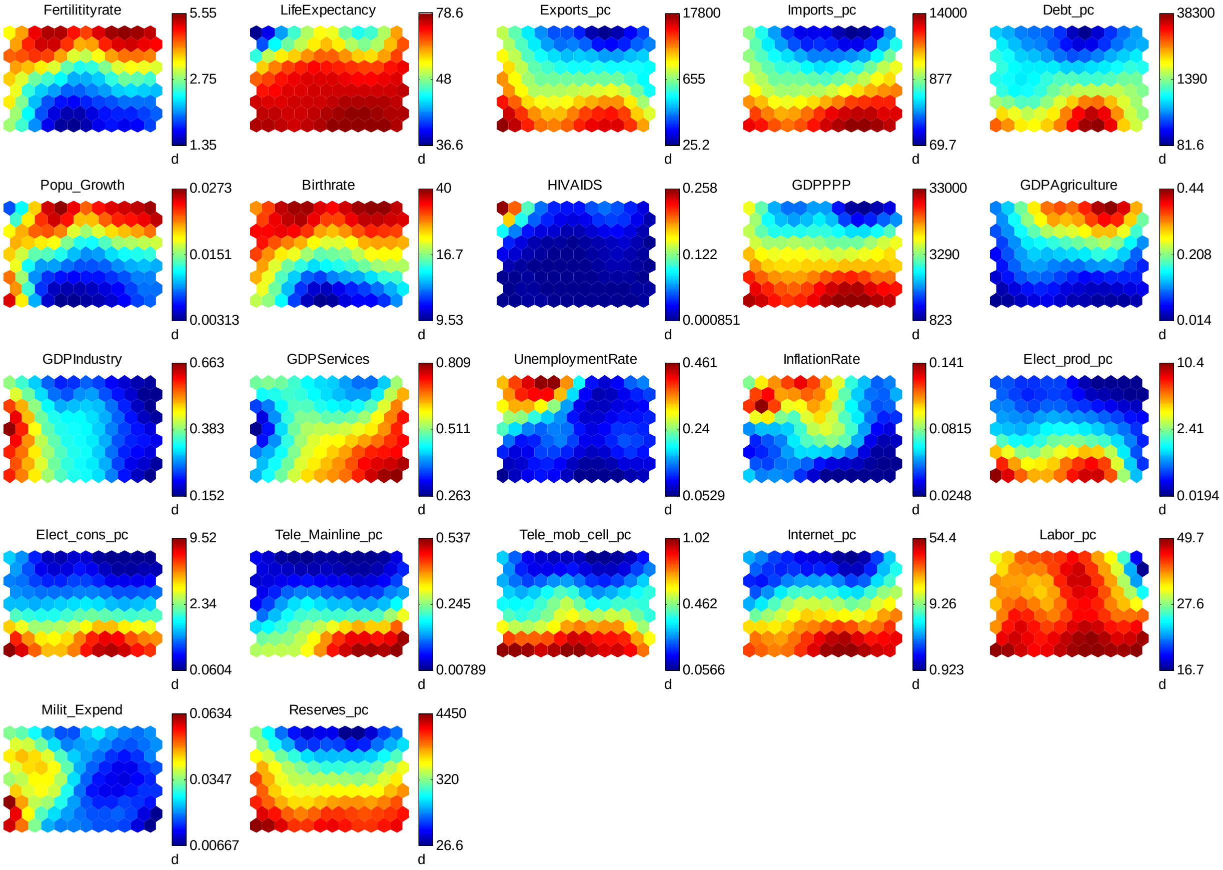


Figure 6. Individual component maps

Most component maps indicate that the properties of SOM are satisfied; the colors change gradually from shades of red indicating high values to shades of blue indicating low values as one moves across the map, vertically, horizontally or diagonally (Figure 6).

One map that displays some problems with this property (gradual transition of colors) is the component map for the variable Population Growth Rate (Figure 8a). One can observe that there is a series of red hexagons in three corners of the map. This could be due to outlier effects related to several relatively small oil producing countries such as Kuwait, Qatar and United Arab Emirates. These countries are located in the bottom left corner on the U-matrix. These countries have high population growth rates, but they do not fit the profile of most countries with a high population growth rate (see description of clusters #2 and #3 in the section below). Most countries with high population growth rates are relatively poor with a low level of economic development and the three above mentioned countries do not fit that profile.

Cluster Analysis

We now discuss how to extract clusters from the U Matrix and we note several clearly defined clusters (see figure 9). Appendix 2 provides a key to the country codes in Figures 5 and 9.

Our discussion proposes 9 clusters, but there is certain amount of arbitrariness in the selection of a number of clusters from a U-matrix. One might consider merging clusters 2 and 3, clusters 7 and 8, and clusters 8 and 9; these merges would yield a six-cluster solution close to that in Figure 4.

The first cluster can be defined as the cluster of countries with the highest estimated rate of HIV infection. The cluster is located in the left top corner of the map. The estimated rates of HIV infection range from 38.8% in Swaziland to 21.3% in Namibia. Not surprisingly, this cluster is characterized by the lowest estimated life expectancy at birth: from 32.23 years in Swaziland to 50.58 years in Botswana. These countries

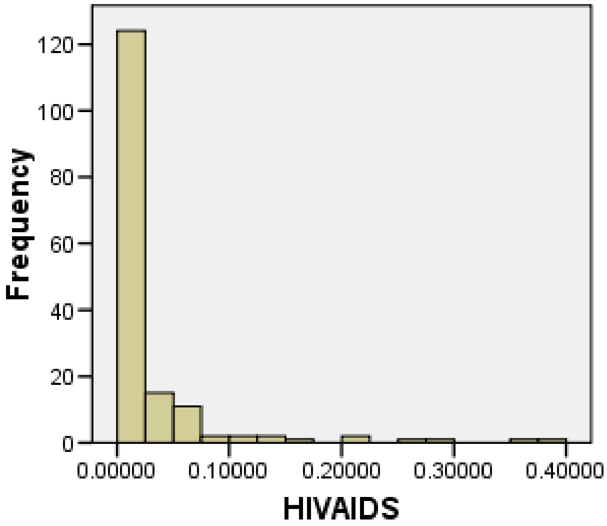


Figure 7(a). HIV/AIDS histogram

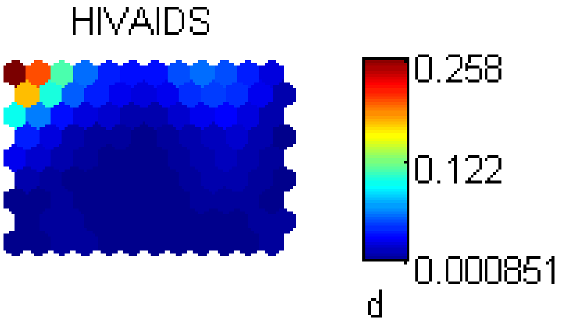


Figure 7(b). HIV/AIDS component map

have a relatively high estimated birth rate but low estimated population growth rate, which can be explained by the extremely high rates of HIV infections and deaths caused by it. This cluster contains countries with a high level of unemployment. Estimated values of most economic variables indicate a low level of economic development.

The second well defined cluster is in the middle of the top part of the map. This cluster primarily consists of African countries but contains some countries from other parts of the world as well. This cluster is characterized by very high estimated unemployment and inflation rates, very high estimated birth rate and population growth rates, and low values of most economic indicators.

The third cluster is located on the top right corner of the map. This cluster also primarily consists of African countries and contains several countries from other parts of the world. This cluster can be described as a cluster of countries with the lowest estimated debt per capita, lowest estimated electricity production per capita, lowest estimated GDP per capita, lowest estimated reserves per capita and the highest estimated rate of agriculture as a percent of GDP.

The fourth cluster is located in the middle of the left side of the map. It can be characterized as a cluster with the highest estimated rate of industry as a percent of GDP and the lowest estimated rate of services as a percent of GDP. These countries have relatively high estimated rates of export per capita, but low estimated values of other economic parameters.

The fifth well defined cluster is on the left bottom corner of the map. This cluster mostly consists of oil producing Middle Eastern countries. They have the highest estimated rate of exports per capita, high estimated rate of industry as a percent of GDP, low estimated unemployment and inflation rates, high estimated GDP per capita. Interestingly, these countries have low values of land line telephone users per capita, but have very high estimated rates of cell phone use.

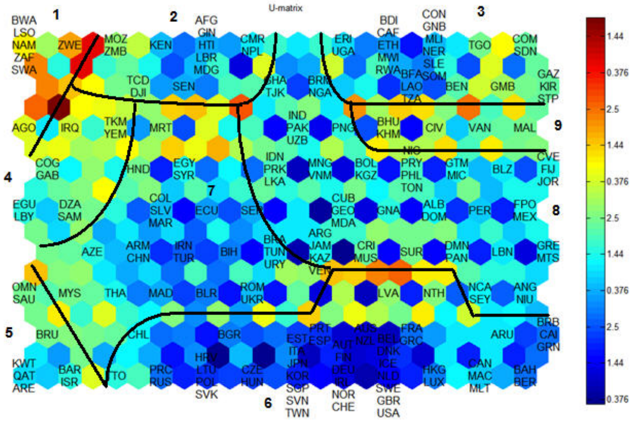


Figure 8. Broad groups identified in the U-matrix

The sixth cluster is located on the bottom part of the map: stretching from its center to the right corner. This cluster consists of Western European countries, the United States, Australia, some Eastern European countries and several other countries from different parts of the world. These countries are characterized by the highest estimated life expectancy, estimated GDP per capita, estimated electricity production and estimated electricity consumption per capita, estimated number of internet and telephone users. These countries have the lowest estimated population growth rate and fertility rate; and low estimated unemployment and inflation rates.

Clusters seven, eight and nine are more homogenous and are more difficult to define. They stretch from the middle part of the map to the middle of its right side. The countries that belong to these clusters generally have medium levels of estimated values describing economic development and standard of living.

In general, most parameters that indicate economic development and quality of life increase from the top of the map to the bottom (see Figure 6). Estimated population growth rate increases from the bottom to the top. Estimated life expectancy steadily decreases as estimated population growth rate increases. Estimated GDP per capita follows the same trend: as population growth rate increases, GDP per capita decreases. GDP Agriculture, GDP Industry and GDP service are split very clearly: the highest percent of Agriculture as a part of GDP is in the first cluster (mostly African countries in top left corner of the map).

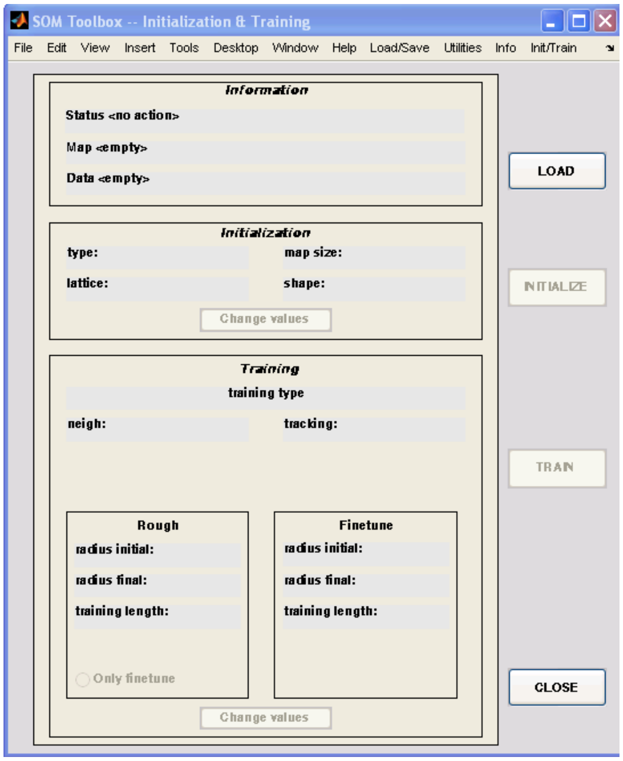
Many countries on the map are organized according to their geographic position, which indicates that many economical and quality of life parameters are characteristic to different parts of the world. However, all of the clusters also include some countries from different parts of the world that are similar in their economic development.

4. Conclusion

In this paper we have demonstrated a case of applying a self organizing Kohonen map algorithm to a multidimensional economical data set that consisted of 22 indicators and 180 countries. We have shown how to apply the algorithm to the data using the Matlab SOM toolbox.

We have identified nine clusters, based on the U-matrix. However, the determination of cluster membership is very subjective, as the algorithm does not assign any cluster membership, but rather shows the color coded distances between estimated vectors of map positions.
The advantage of the self organizing Kohonen map methodology is a visualization of values of multiple parameters on component maps (which exhibit the SOM property for the most part) and a visualization of clustering results on the U-matrix. It provides an easy way for determining the common characteristics of different clusters based on an examination of the component maps.

Appendix 1---Screen shot of the SOM Graphical User Interface



Status <no action>

Map <empty>

Data <empty>

Appendix 2: Country Codes (abbreviation and name)

AFG	Afghanistan
ALB	Albania
DZA	Algeria
ASO	American Samoa
AND	Andorra
AGO	Angola
ANG	Anguilla
ARG	Argentina
ARM	Armenia
ARU	Aruba
AUS	Australia
AUT	Austria
AZE	Azerbaijan
BAH	Bahamas
BAR	Bahrain
BGD	Bangladesh
BRB	Barbados
BLR	Belarus
BEL	Belgium
BLZ	Belize
BEN	Benin
BER	Bermuda
BHU	Bhutan
BOL	Bolivia
BIH	Bosnia and Herzegovina
BWA	Botswana
BRA	Brazil
BVA	British Virgin Is.
BRU	Brunei
BGR	Bulgaria
BFA	Burkina Faso
BRM	Burma
BDI	Burundi
KHM	Cambodia
CMR	Cameroon
CAN	Canada
CVE	Cape Verde
CAI	Cayman Islands
CAF	Central African R.
TCD	Chad
CHL	Chile
CHN	China
COL	Colombia
COM	Comoros
CON	Congo, Democratic Republic
COG	Congo, Republic
CRI	Costa Rica
CIV	Cote d'Ivoire
HRV	Croatia
CUB	Cuba
CYP	Cyprus
CZE	Czech Republic
DNK	Denmark

DJI	Djibouti
DMN	Dominica
DOM	Dominican Republic
ECU	Ecuador
EGY	Egypt
SLV	El Salvador
EGU	Equatorial Guinea
ERI	Eritrea
EST	Estonia
ETH	Ethiopia
FIJ	Fiji
FIN	Finland
FRA	France
FPO	French Polynesia
GAB	Gabon
GMB	Gambia, The
GAZ	Gaza Strip
GEO	Georgia
DEU	Germany
GHA	Ghana
GIB	Gibraltar
GRC	Greece
GRN	Greenland
GRE	Grenada
GUA	Guam
GTM	Guatemala
GRN	Guernsey
GIN	Guinea
GNB	Guinea-Bissau
GNA	Guyana
HTI	Haiti
HND	Honduras
HKG	Hong Kong
HUN	Hungary
ICE	Iceland
IND	India
IDN	Indonesia
IRN	Iran
IRQ	Iraq
IRL	Ireland
ISR	Israel
ITA	Italy
JAM	Jamaica
JPN	Japan
JOR	Jordan
KAZ	Kazakhstan
KEN	Kenya
KIR	Kiribati
PRK	Korea, North
KOR	Korea, South
KWT	Kuwait
KGZ	Kyrgyzstan
LAO	Laos
LVA	Latvia

LBN	Lebanon
LSO	Lesotho
LBR	Liberia
LBY	Libya
LIE	Liechtenstein
LTU	Lithuania
LUX	Luxembourg
MAC	Macau
MAD	Macedonia
MDG	Madagascar
MWI	Malawi
MYS	Malaysia
MAL	Maldives
MLI	Mali
MLT	Malta
MRT	Mauritania
MUS	Mauritius
MTT	Mayotte
MEX	Mexico
MIC	Micronesia, Federated States
MDA	Moldova
MON	Monaco
MNG	Mongolia
MGR	Montenegro
MTS	Montserrat
MAR	Morocco
MOZ	Mozambique
NAM	Namibia
NAU	Nauru
NPL	Nepal
NLD	Netherlands
NTH	Netherlands Antilles
NCA	New Caledonia
NZL	New Zealand
NIC	Nicaragua
NER	Niger
NGA	Nigeria
NIU	Niue
NOR	Norway
OMN	Oman
PAK	Pakistan
PAL	Palau
PAN	Panama
PNG	Papua New Guinea
PRY	Paraguay
PER	Peru
PHL	Philippines
POL	Poland
PRT	Portugal
PRC	Puerto Rico
QAT	Qatar
ROM	Romania

RUS	Russia
RWA	Rwanda
SAM	Samoa
SMR	San Marino
STP	Sao Tome/Principe
SAU	Saudi Arabia
SEN	Senegal
SER	Serbia
SEY	Seychelles
SLE	Sierra Leone
SGP	Singapore
SVK	Slovakia
SVN	Slovenia
SOM	Somalia
ZAF	South Africa
ESP	Spain
LKA	Sri Lanka
SDN	Sudan
SUR	Surinam
SWA	Swaziland
SWE	Sweden
CHE	Switzerland
SYR	Syria
TWN	Taiwan
TJK	Tajikistan
TZA	Tanzania
THA	Thailand
TIM	Timor-Leste
TGO	Togo
TON	Tonga
TTO	Trinidad and Tobago
TUN	Tunisia
TUR	Turkey
TKM	Turkmenistan
TUV	Tuvalu
UGA	Uganda
UKR	Ukraine
ARE	United Arab Emirates
GBR	United Kingdom
USA	United States
URY	Uruguay
UZB	Uzbekistan
VAN	Vanuatu
VEN	Venezuela
VNM	Vietnam
WNK	West Bank
YEM	Yemen
ZMB	Zambia
ZWE	Zimbabwe

Appendix 3: Matlab Source Code

```
% Exploratory Analysis of 2007 CIA
Factbook Data Using Self Organizing
Maps
% KOM Group
% Bentley University
% April 20, 2009

% Make the data
%sD=som_read_data('final_data.dat','x')
sD=som_read_data('data.dat','x')

%Normalize the data
sD=som_normalize(sD,'log')
sD=som_normalize(sD,'var')

% Make the SOM
sM=som_make(sD, 'msize', [9 12],
'lattice', 'hexa');
% seq training algorithm
% sM=som_make(sD, 'algorithm','seq');

sMlab=som_autolabel(sM,sD)
U=som_umat(sM);
Um=U(1:2:size(U,1),1:2:size(U,2));

% Basic visulization

% Figure(1)--D-matrix
figure(1)
h=som_cplane(sM, Um(:));
set(h,'Edgecolor','none'); title('D-
matrix')

figure(2)
[Pd,V,me,l] = pcaproj(sD,2); Pm =
pcaproj(sM,V,me); % PC-projection
Code = som_colorcode(Pm); % color
coding
hits = som_hits(sM,sD);  % hits
U = som_umat(sM); % U-matrix
Dm = U(1:2:size(U,1),1:2:size(U,2)); %
distance matrix
Dm = 1-Dm(:)/max(Dm(:));
Dm(find(hits==0)) = 0; % clustering
info

som_cplane(sM,Code,Dm);
hold on
som_grid(sM,'Label',cellstr(int2str(hi
ts)),...

'Line','none','Marker','none','Labelco
lor','k');
hold off
```

```
%Figure(3) shows PCA and initial
clusters
figure(3)
[Pd,V,me,l] = pcaproj(sD,2);
Pm = pcaproj(sM,V,me); % PC-projection
Code = som_colorcode(Pm); % color
coding
hits = som_hits(sM,sD);  % hits
U = som_umat(sM); % U-matrix
Dm = U(1:2:size(U,1),1:2:size(U,2)); %
distance matrix
Dm = 1-Dm(:)/max(Dm(:));
Dm(find(hits==0)) = 0; % clustering
info

subplot(1,3,1)
[c,p,err,ind] = kmeans_clusters(sM,
7); % find at most 7 clusters
plot(1:length(ind),ind,'x-')
[dummy,i] = min(ind)
cl = p{i};

subplot(1,3,2)
som_cplane(sM,Code,Dm)

subplot(1,3,3)
som_cplane(sM,cl)

%Figure(4) shows Davies-Bouldin index
and K-means clusters
figure(4)
som_grid(sM,'Coord',Pm,'MarkerColor',C
ode,'Linecolor','k');
hold on, plot(Pd(:,1),Pd(:,2),'k+'),
hold off, axis tight, axis equal
title('PC projection')

%Figure(5)--U-matrix
figure(5)
som_show(sMlab,'umat','all')
som_show_add('label',sMlab,'Textsize',
12)

%Figure(6)--Complane
figure(6)
som_show(sM,'comp','all')

U=som_umat(sM);
Um=U(1:2:size(U,1),1:2:size(U,2));

% Qe and Te
[qe,te]=som_quality(sM,sD)
```

REFERENCES:

Davies, D. L. and D. W. Bouldin. 1979. "A cluster separation measure", *IEEE Trans. on Pattern Analysis and Machine Intelligence* **PAMI-1(2):** 224-227.

Deichmann, J., Eshghi, A., Haughton, D., Sayek, S. and Woolford, S. 2007. “Measuring the International Digital Divide: An application of Kohonen Self Organizing Maps”, *International Journal of Knowledge and Learning*, 3(3): 552:575.

Ingaramo, D.A., Leguizamón, G. and Errecalde, M. 2005. “Adaptive clustering with artificial ants”, *Journal of Computer Science and Technology*, 5(4):264-271.

Larose, D.T. 2005. “Kohonen Networks”, *Discovering Knowledge in Data: An Introduction to Data Mining*: Wiley & Sons, Inc.

Nguyen, P. Haughton, D. and I. Hudson. 2008. “Living Standards of Vietnamese provinces: a Kohonen map”, *Case Studies in Business, Industry and Government Statistics,* **2(2)**, pp. 109-113.

Peeters, L. 2006. Interactive comment on “Exploratory data analysis and clustering of multivariate spatial hydrogeological data by means of GEO3DSOM, a variant of Kohonen’s Self-Organizing Map” by L. Peeters et al.

"Synapse." from http://www.peltarion.com/WebDoc/index.html.

Vesanto, J. 2005. “SOM implementation in SOM Toolbox”.

Vesanto, J., Himberg, J., Alhoniemi, E. and Parhankangas, J. 1999. “Self-organizing map in Matlab: the SOM Toolbox”, *Proceedings of the Matlab DSP Conference*. Espoo, Finland, 35-40.

Vesanto, J. 2005. SOM algorithm implementation in SOM Toolbox. Hydrology and Earth System Sciences Discussions.

Correspondence: ghua@bentley.edu