2012 16th International Conference on Information Visualisation

A New Automated Hierarchical Clustering Algorithm Based On Emergent Self Organizing Maps

Seyed Vahid Moosavi, Qin Rongjun
Future Cities Laboratory, Department of Architecture, ETH Zurich, 8092 Zurich, Switzerland
svm@arch.ethz.ch, rqin@student.ethz.ch

***Abstract*— Emergent Self Organizing Map (ESOM) has been shown as a powerful nonlinear data transformation and visualization method. In [13] based on ESOM and some of its derivatives, U-Matrix and P-Matrix, a powerful automated clustering algorithm, U*C, is introduced, and it is shown that the algorithm performs better than the some of the benchmark algorithms. However, the mentioned algorithm is suitable for partitional clustering tasks, while in most of the real cases, because of the nature of the data sets (not the ESOM training algorithm) a hierarchical structure in the data can be assumed. In this paper, based on the main ideas of U*C algorithm and underlying meaning of the U-Matrix, we introduced an automated hierarchical clustering algorithm, which performs well for real data sets. After testing with some test cases, we applied the proposed algorithm on a real data set, including different energy, ICT and Urban related indicators of European and central Asian countries. The proposed algorithm identified the hierarchical groups among the selected countries.**

Keywords- Emergent Self Organizing Map, Hierarchical Clustering, Visualization, Automated Clustering

I. INTRODUCTION

Self Organizing Map (SOM) as a powerful nonlinear data transformation method has been applied successfully in thousands of application domains [16]. One of the main applications of SOMs are in data clustering, in which the problem is finding similar groupings of the objects based on a set of selected features, a (dis)-similarity measure and without any prior knowledge about the possible groupings. In classic SOM clustering algorithms, usually after training the SOM, each neuron is representing a group of data and then, each neuron can be considered as a data cluster or after SOM training, different clustering algorithms can be applied on neurons instead of original data. [15] However, in large data sets usually small SOMs with small number of neurons (as the number of clusters) cannot perform well.

In [14] the idea of Emergent Self Organizing Map (ESOM), with many (usually several thousands) of neurons is introduced. Based on [12]: "*This leads to the definition of ESOM as SOM producing a nontrivial U-Matrix on which the terms "watershed" and "catchment basin" are meaningful and which are cluster conform.*" Using ESOM and U-Matrix for cluster visualization of a high dimensional data space, has several advantages over existing famous algorithms (e.g. K-means, single linkage or Ward algorithm):

- In ESOM there is no need to a prior knowledge for deciding on the number of clusters. [6,13]
- The cluster shapes can be in any form, while for example in K-means with Euclidean distance, the algorithm leads just to clusters with super- spherical shapes. [13]
- The final clusters are visualized in a two (three) dimensional gird (U-matrix) in a very attractive manner, in which most of non-expert stakeholders of the subject can easily identify the clusters or discuss them.

The U*C algorithm as a powerful automated clustering algorithm on top of U-Matrix and P-Matrix (Based on the density of similar data around each neuron) of ESOM is introduced and it has been shown that its performance on several hard clustering problems is superior to the existing algorithms such as Single Linkage, Ward and K-Means. [13]

Further, it has been shown that comparing to ESOM clustering, usually Particle Swarm Based Clustering (PSBC) methods have poor results in terms of topographic mapping and cluster formation. However they can be seen in a same class of clustering algorithms. [3]

In practical applications, ESOM clustering has been applied in different domains including Urbanism [1], Music clustering [10], Stock Market Analysis [11], domestic violence detection [9], and text mining and document analysis [2,8].

However, it should be noted that although the developed algorithm for automatic clustering of ESOM (U*C) is performing well, it is not designed to identify the hierarchical orders in the clusters; while in most of real cases there are natural hierarchical orders inside each cluster, which are visible in the U-maps.

In this paper, after some discussions on U*C and its similar algorithms, we present a new recursive procedure for finding the hierarchical clusters within each cluster, which is based on an iterative call of a well-known image segmentation algorithm and a threshold value. In the last section, after experimenting with some well known data set, we present the result of applying the proposed algorithm on a data set, related to the countries in Europe and Central Asia, consisting several energy, ICT and urban related indicators.

II. THE DESCRIPTION OF THE PROPOSED HIERARCHICAL CLUSTERING ALGORITHM

The U*C algorithm [13] consists of two main parts:

- *Immersion*

1550-6037/12 $26.00 © 2012 IEEE
DOI 10.1109/IV.2012.52

- *Cluster Assignment*

Immersion step itself includes two main parts: a gradient descent move from each point, *x,* in U-matrix until reaching to a local minimum, *y*, which is probably inside the cluster area, comparing to the starting point *x*, and sequentially starting a gradient ascent from *y* on P-Matrix until reaching to a local maximum, *I*, which is called *immersion* point for *x*.

The rationale behind finding immersion point is that by this method we can make sure that in immersion points we are probably inside the cluster of the starting point without passing the border of any cluster. This idea is true because usually cluster centers are in the local minimum points in U-matrix and for those areas of the map with low density and low U-values (but not minimum), we should find the cluster centers in local maximum of P-matrix. Note that, if *y* is on the border of the ESOM map, then probably it has a low P-value and then with the next gradient ascent we will move to the center of the cluster.

In the second step of the U*C algorithm, *Cluster Assignment,* after finding the immersion points (as representatives for the possible clusters), we assume that each point and its immersion point are in the same cluster and if we partition the ESOM map based on these immersion points, we can find the natural emergent cluster areas.

Then in this step of the U*C algorithm, U* matrix (which is a combination of U and P matrices [13]) will be partitioned, based on the identified immersion points in the previous step, using any *Watershed detection* algorithm like the well known method, presented in [4]. The pseudo code of the described procedure is as follows [13]:

U*C clustering Algorithm: given U-Matrix, P-Matrix, U*-Matrix, I = {};

Immersion:

For all positions n of the grid:

1) From position n follow a gradient descent on the U-Matrix until a minimum is reached in position u

2) From position u follow a gradient ascent on the P-Matrix until a maximum is reached in position p.

3) I = I {p}; Immersion(n) = p.

Cluster assignment:

1) Calculate the watersheds for the U*-Matrix (e.g. using Luc/Soille (1991)).

2) Partition I using these watersheds into clusters $C_1,\ldots C_c$

3) Assign a data point x to a cluster C_j if Immersion(bm(x))$\in$ _C_j.

Based on our experience with the real data sets, using watershed algorithms, sometimes these algorithms like the algorithm, presented in [4], in their original forms find a lot of small watersheds, which are clearly more than those interested visual clusters in the ESOM map. For example, in figure 1, on the left side, the visual clusters of a sample high dimensional data in U-matrix is shown and of course considering the different values (U-heights) in the border of watersheds, one can find different number of clusters with different resolutions. In other words, selecting the number of clusters is not easy and depends on a threshold value for selecting the edges of the clusters and as it is shown in Fig. 1(right figure) from the result of applying that watershed algorithm, which is described in [4], in Matlab software environment the total number of identified watersheds is much more than what is expected based on the visualization of the clusters on the left side of Fig. 1.

For the issue of validation, it is important to note that the possible results of this algorithm are just based on the topographic mapping of ESOM algorithm and the identified hierarchies can be validated visually by checking the original map of ESOM after training.

However, by changing the number of units for neighborhood search and changing the threshold in the required difference between two neighborhood values for being in the edge point, it is possible to reach to different results. Then, for example in Fig. 1 left side if we put a threshold on the values of the matrix (e.g. U-heights in U-matrix) and use a simple rule for filtering those values less than that threshold, we will find different clustering results. Some possible results on this sample data are shown in Fig. 2.

Here, the problem of selecting the threshold value for finding the clusters is the same as the problem of selecting the number of clusters as a prior knowledge in other clustering algorithms like k-means. And although there are some criteria in the literature for optimizing the number of clusters (e.g. Elbow point method), in real cases it can be up to the final stakeholders and decision makers to identify different clusters based on their discussions and final consensus.

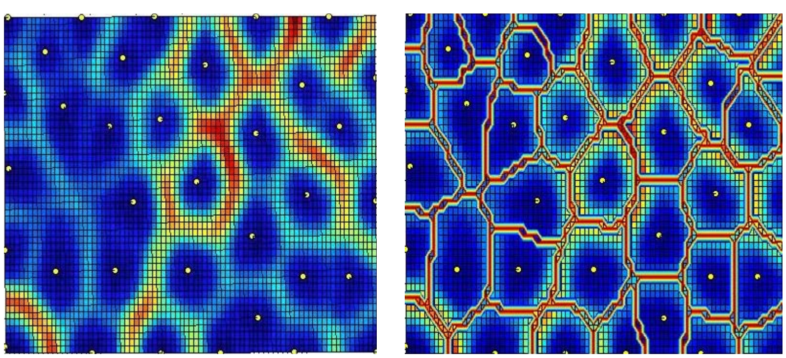


Figure 1-The result of applying the watershed algorithm on a sample U-Matrix

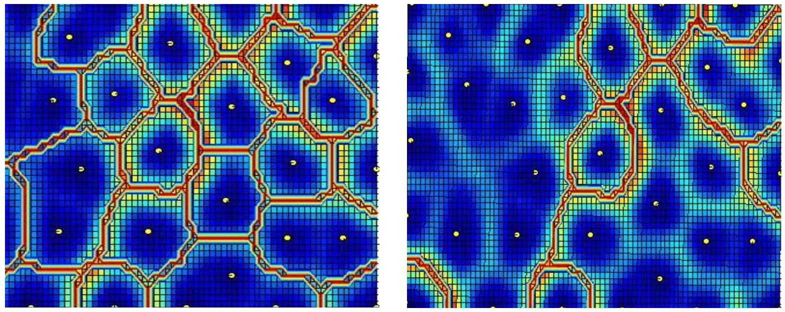


Figure 2- different clusters based on different threshold for edge detection; with smaller values we have more clusters

In addition to U*C clustering algorithm, there are other similar algorithms [6,7], in which there is no need to do the immersion step in U*C and just based on the method of Region growing [5] from the domain of image processing, the cluster segments on U-matrix or U*matrix can be identified, and it is just required to select an appropriate threshold value the region growing algorithm. For example, the U*F algorithm [6] works very well on the same data sets that have been analyzed in [13]. But another possible way for dealing with the problem of finding a "good" segmentation (which is equal to the number of clusters) on ESOM map is to assume a hierarchical structure for the clusters. In other words, we can assume that those borders with higher values are the borders of higher level clusters and those borders with lower values are the borders of the sub-clusters within the higher level clusters. In fact, with this point of view to U-matrix and its underlying meaning, in existing algorithm like U*C and U*F, we are loosing some valuable information of hierarchical groups within the data.

Therefore, if one starts with a good threshold value to find the first level clusters, then by repeating the map segmentation procedure within each cluster, with new scaled-up U-values between 0 and 1, finally a natural hierarchical cluster can be found. The described idea can be implemented as an iterative process on top of U-matrix (or U*-matrix) as follows:

0-Points for the clustering <--------- all of the points (neurons) in ESOM gird

1-Scale up the values to [0 1]

2-Find the first clusters and their edges using region growing algorithm, based on the selected threshold value. (This value can be decided, based on the visual results of clustering on ESOM final map and agreement among the stakeholders.)

3-For all of the identified clusters do:

3-1-Points for clustering <--------- all of the points (neurons) within the cluster

3-2-Repeat line 1,2, 3 for each sub-cluster until max_value- min_value > Threshold for the selected group of nodes in that (sub-)cluster.

4-assign hierarchical cluster labels to the original data, based on the labels of their corresponding BMUs in the trained ESOM.

As it can be seen from the above procedure, in this algorithm unlike U*C algorithm, there is no separate step to find the immersion points, and in fact it is done within watershed detection step.

Further, in this paper instead of using the presented algorithm in [4] for watershed detection, we used a region growing algorithm similar to those used algorithm in [6, 7], which starts from every point in the map and if based on the selected threshold value, its value is in the same range with its neighbor neurons, it has the same label (cluster name) with its neighbor neurons. And if its value comparing to its neighbor's values are greater than the threshold, then it is on the edge of the cluster. This simple procedure can be applied either in U-matrix or U*-matrix. [13]

Another important point is that in this way, if there is no evident hierarchy in the clusters, then the final result is just the same as the result of one level clustering algorithms like U*C and U*F.

The result of applying the described procedure on the above example is shown in figure 3.

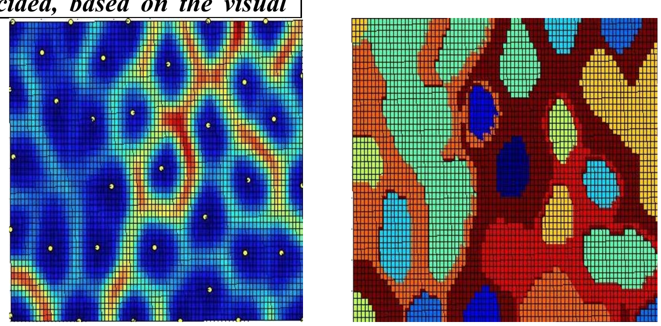


Figure 3- The identified hierarchical clusters in U-Matrix

For the issue of validation, it is important to note that the possible results of this algorithm are just based on the topographic mapping of ESOM algorithm and the identified hierarchies can be validated visually by checking the original map of ESOM after training.

III. Some Experimental Results

In this section, we present the result of applying the proposed algorithm to some of the data sets from Fundamental Clustering Problem Suite (FCPS), which have been used in [6,13]. As it can be seen in figure 4, the proposed procedure for finding clusters works well for these data sets and since there is no hierarchical order in the data set, the final clusters are in one level.

In all of experiments in this paper, we used the Databionic open source software[1] for training ESOM and the proposed hierarchical algorithm is done in Matlab software environment.

In the above examples, there is no hierarchical structure in the data sets, but it shows the performance of the region growing algorithm, with the selected threshold value, for map segmentation.

Then, as it is mentioned before, the idea is to apply the same cluster detection algorithm in a recursive way within each cluster to find the hierarchical clustering orders. In the next section, we applied the proposed algorithm to a real data set to find the hierarchical clusters within the data.

IV. Hierarchical Clustering Of Different Countries Based On A Set Of Urban, Energy And High-tech Related Factors

In this section, we applied the proposed hierarchical clustering algorithm to a real data set from the World Bank Data Base[2]. We selected the three following groups of indicators of European and Central Asian countries:

- Density and Urbanization
- Electricity and Energy
- High-tech and ICT

These features are selected since they are among the main important features for presenting a country as a complex phenomenon, but it should be noted that these features are selected arbitrarily and not based on a specific domain knowledge, and the main idea here, is to show the possible applications of presented hierarchical clustering algorithm for real world application domains.

Each of these groups of indicators consists of several sub-indicators (features) as depicted in table1.

Table 1- selected features for representation of the countries

No.	Main indicator group	Sub-indicator
1	Density and Urbanization	1. Population-density-(people-per-sq-km-of-land area) 2. Population-in-the-largest-city-(%-of-urban-population) 3. Urban-population-(%-of-total)
2	Electricity and Energy	4. Alternative-and-nuclear energy-% 5. Electric-power-consumption-kWh-per-capita 6. Electric-power-losses-% 7. Electricity-production-from-hydroelectric-sources-% 8. Electricity-production-from-nuclear-sources-% 9. Electricity-production-from-oil-gas-and-coal-sources-% 10. Energy-imports-net-% 11. Fossil-fuel-energy-consumption-%
3	High-tech and ICT	12. High-technology-exports-(%of-manufactured-exports 13. ICT-goods-exports-(%of-total-goods-exports) 14. Internet-users-(per-100-people) 15. Mobile-cellular-subscriptions-(per100 people) 16. Telephone-lines-(per-100-people)

As it is shown in Table 1 there are totally 16 indicators, making a 16 dimensional data space and like any other data analysis procedure, it could be better if we apply some feature selection or feature extraction algorithms (e.g. PCA or ICA) for reducing the number of dimensions, but in this example we just applied the ESOM algorithm on the normalized data set. Based on the availability of the data, we selected finally 42 countries with their related data for year 2008. (Table 2)

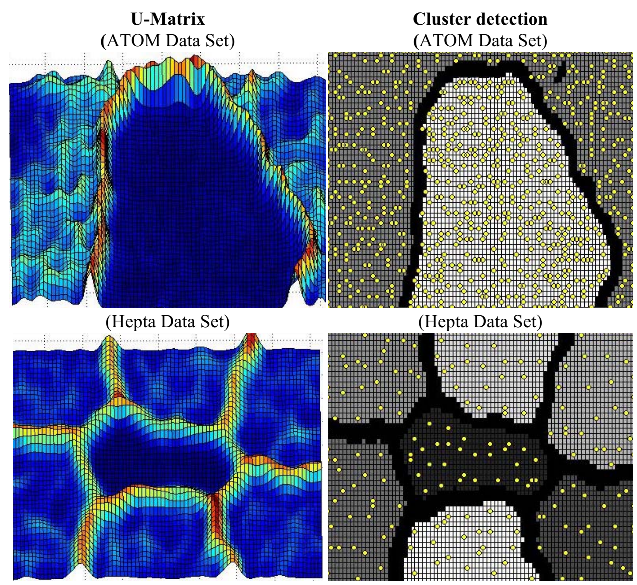
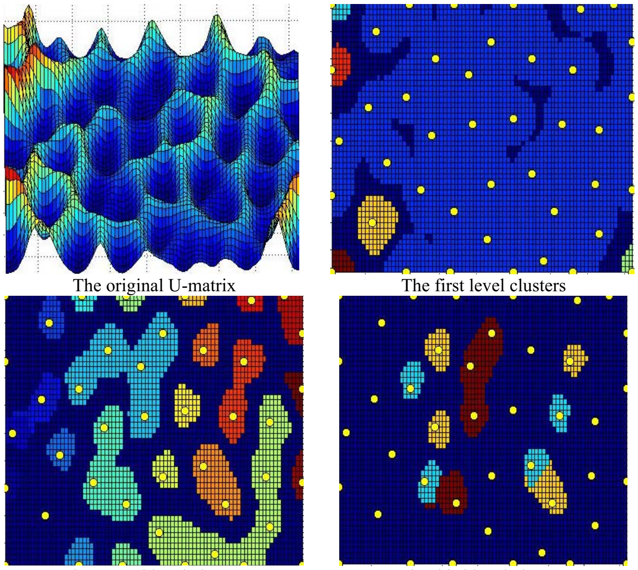


Figure 4-The identified clusters in the U-Matrices of three benchmark data

[1] http://databionic-esom.sourceforge.net/

[2] http://data.worldbank.org/



The original U-matrix — The first level clusters

The second level clusters — The third level clusters

Figure 5- The identified hierarchical clusters in different levels of U-matrix of the selected countries. Dark areas in each level do not belong to any cluster.

Table 2-Selected countries for the analysis

No .	Name	No.	Name	N o.	Name	N o.	Name
1	Albania	12	Denma rk	23	Kazakhstan	34	Russian Federation
2	Armenia	13	Estonia	24	Kyrgyz Republic	35	Slovenia
3	Austria	14	Finland	25	Latvia	36	Spain
4	Azerbaijan	15	France	26	Lithuania	37	Sweden
5	Belarus	16	Georgi a	27	Luxembourg	38	Switzerland
6	Belgium	17	Germa ny	28	Moldova	39	Turkey
7	Bosnia and Herzegovina	18	Greece	29	Netherlands	40	Ukraine
8	Bulgaria	19	Hungar y	30	Norway	41	United Kingdom
9	Croatia	20	Iceland	31	Poland	42	Uzbekistan
10	Cyprus	21	Ireland	32	Portugal		
11	Czech Republic	22	Italy	33	Romania		

The results of applying ESOM and the proposed hierarchical clustering algorithm are depicted in figure 5. In this case, we finally reached to three levels of hierarchy. As it can be seen in the original U-matrix, (top left in figure 5), in small parts around the U-matrix have higher edges comparing to those cluster areas in the center of the matrix. Therefore, in the first level of the clustering algorithm (Top right in Fig. 5), the main area of the map is in one cluster, while we have 5 small distinct clusters around the map. The next two figures show the main lower level clusters within the main cluster in the first level, while the other areas (in first level) do not consist any more hierarchies. Note that those areas of the maps with dark color have no cluster labels.

Based on the identified three level clusters in the ESOM map, the labels of the countries based on their corresponding BMUS (yellow points in figure 5) have been selected. In Fig. 6 the countries and their corresponding cluster numbers have been shown in a hierarchical form.

Normally, after the clustering step in a standard data analysis process, we need to further analyze the identified clusters, based on their internal characteristics for knowledge creation and the next steps in data mining process, but in this work since the focus of the paper was just on the presentation of the hierarchical clustering algorithm on ESOM, we stopped in this step.

V. CONCLUSION

Emergent Self-Organizing Map has been shown as a powerful nonlinear visualization method for high dimensional data spaces. Based on some characteristics of ESOMs, it is shown that the emergent clusters within data can be visualized on U-matrix. In U*C algorithm [13] a powerful automated method for identifying the visualized clusters is introduced. However, in real cases, the number of visualized clusters on the map is not determined and clear and sometimes, a hierarchy of clusters can be assumed on the final ESOM maps. But the existing automated algorithms [6,13] cannot find the hierarchical orders among the clusters. In this paper, based on the main ideas of U*C algorithm and the underlying meaning of U-matrix, we proposed a simple recursive method to identify the visualized clusters in a hierarchical order. Finally, we applied the proposed algorithm for grouping of a set of European and Central Asian countries, based on 16 selected features of these countries. The final result is interesting.

VI. ACKNOWLEDGMENTS

This work was established at the Singapore-ETH Centre for Global Environmental Sustainability (SEC), co-funded by the Singapore National Research Foundation (NRF) and ETH Zurich.

VII. REFERENCES

[1] Behnisch, M., Ultsch, A.: "Urban data mining: spatiotemporal exploration of multidimensional data," *Building Research & Information ,2009, Vol. 37, Nr. 5-6.*

[2] De Mazière, P.A., and Van Hulle, M.M. "A Clustering Study of a 7,000 EU Document Inventory using MDS and SOM," Expert Systems with Applications, 2011, 38, 8835-8849.

[3] Herrmann, L., Ultsch, A., "Clustering with Swarm Algorithms Compared to Emergent SOM," In Advances in Self-Organizing Maps,7th International Workshop, WSOM 2009, St. Augustine, Florida, 2009.

[4] Luc, V., Soille, P., "Watersheds in Digital Spaces: An Efficient Algorithm Based on Im-mersion Simulations," IEEE Transactions of Pattern Analysis and Machine Intelligence, 1991, Vol. 13(6), 583-598.

[5] Mancas, M., Gosselin, B., and Macq, B., “Segmentation using a region-growing thresholding,” Proc. SPIE 5672, 388 2005.

[6] Moutarde, F., Ultsch, A. “U*F Clustering: a new performant Cluster-mining method based on segmentation of Self-Organizing Maps,” Proc. Workshop on Self-Organizing Maps (WSOM 2005), Paris, France, 2005, pp. 25-32

[7] Opolon, D. , Moutarde, F., “Fast semi-automatic segmentation algorithm for Self Organizing Maps,” Proc. of ESANN'2004, Bruges, 28-30 avril 2004, p. 507-512.

[8] Poelmans, J., Van Hulle, M.M., Viaene, S., Elzinga, P., and Dedene, G., “Text Mining with Emergent Self Organizing Maps and Multi-Dimensional Scaling: A comparative study on domestic violence,” Applied Soft Computing, 2011, 11(4), 3870-3876.

[9] Poelmans, J, Elzinga, P., Viaene, S., Van Hulle, M.M.,& Dedene, G., “How Emergent Self Organizing Maps can help counter domestic violence,” Proc. World Congress on Computer Science and Information Engineering (CSIE 2009) (Los Angeles/Anaheim, USA), March 31 - April 2, 2009), 4, pp. 126-136.

[10] Risi, S., Mörchen, F., Ultsch, A., Lewark, P. , “Visual mining in music collections with Emergent SOM,” Proc. Workshop on Self-Organizing Maps (WSOM '07), Bielefeld, Germany, 2007, ISBN: 978-3-00-022473-7

[11] Ultsch, A., Locarek-Junge, H.: “Knowledge Discovery in Stock Market Data,” In Locarek-Junge, H. et al. (Eds.) Classification as a Tool for Research, Proc. 11th IFCS Biennial Conference, 2009.

[12] Ultsch, A., “Emergence in Self-Organizing Feature Maps,” Proc. Workshop on Self-Organizing Maps (WSOM '07), Bielefeld, Germany, 2007, ISBN: 978-3-00-022473-7

[13] Ultsch, A., “Clustering with SOM: U*C,” Proc. Workshop on Self-Organizing Maps (WSOM 2005), Paris, France, 2005, pp. 75-82

[14] Ultsch, A. “Data Mining and Knowledge Discovery with Emergent Self-Organizing Feature Maps for Multivariate Time Series,” Kohonen Maps, 1999, pp. 33-46

[15] Vesanto, J., Alhoniemi, E., “Clustering of the Self-Organizing Map,” IEEE Transactions on Neural Networks, May 2000, 11(3):586-600.

[16] Van Hulle M.M., “Self-Organizing Maps,” Handbook of Natural Computing: Theory, Experiments, and Applications, G. Rozenberg, T. Baeck, J. Kok (eds.), Springer, 2011.

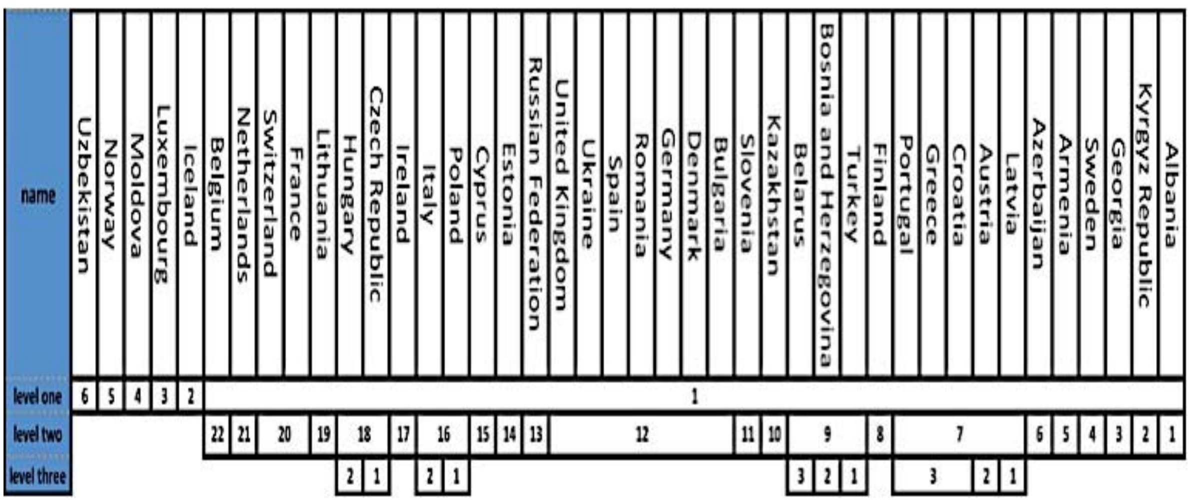


Figure 6- Hierarchical groups within European and Central Asian Countries, based on the selected sample features