Cluster Analysis of the World Gross-Domestic Product Based on the Emergent Self-Organization of a Swarm

Michael C. Thrun[1]

Abstract

Cluster analysis is a task of unsupervised classification which seeks high-dimensional structures if natural clusters exist in data. The Databionic swarm (DBS) can adapt itself to structures of natural clusters characterized by distance and density based structures resulting in a topographic map and clustering. It is the first swarm-based technique that shows emergent properties while exploiting concepts of swarm intelligence, self-organization, and game theory. DBS was applied to the World GDP dataset which was constructed by selecting the purchasing power parity converted gross domestic product (GDP) per capita for the years: 1970-2010. The dynamic time warping distances were used to compute the optimal alignment between time series. The number of clusters was derived from, and the quality of the clustering and were verified by the topographic map which is a 3D representation of data structures. A clear cluster structure is also shown in heat map and silhouette plot. The rules deduced from CART show that the clusters are defined by an event occurring 2001. In its aftermath, the world economy was experiencing its first synchronized global recession in a quarter-century. Therefore, the first cluster consists not-affected, mostly African and Asian countries and a second cluster consists of affected countries which are mostly European and American countries.

Keywords: *machine learning, cluster analysis, swarm intelligence, visualization, self-organization*
JEL Classification: O47, F01, C380
DOI: 10.14659/SEMF.2018.01.53

1 An Approach to Cluster Analysis for Multivariate Time Series

The multivariate time series inspected in this work covers repeated measures of the gross domestic product (GDP) of 190 countries published in (Heston et al., 2012). The GDP consists of the total market value of all final goods and services produced in a country. Thus, the GDP is an indicator of economic achievement of a country Mazumdar (2000). Each country's data has to be converted into common currency to compare the GDP between various currencies. An exchange rate is defined through the purchasing power parity PPP at which the currency of a country is converted into that of another country to purchase the same volume of goods and services in both countries (Rogoff, 1996).

The World GDP data set of was extracted from the multivariate time series of (Heston et al., 2012) by selecting the PPP-converted GDP per capita for the years from 1970 to 2010 by

[1] Corresponding author: University of Marburg, Data Bionics Research Group, Hans-Meerwein-Straße 6, D-35032 Marburg, mthrun@mathematik.uni-marburg.de.

Leister (2016). In this work, the World-GDP data set will be investigated in the context of economic similarity between nations by using cluster analysis.

In cluster analysis, the methods rely on some concept of the similarity between pieces of information encoded in the data of interest. However, no accepted definition of clusters exists in the literature (Hennig et al., 2015, p. 705). Additionally, Kleinberg showed for a set of three simple axioms, scale-invariance, consistency, and richness, that there exists no clustering algorithm which can satisfy all three (Kleinberg, 2003). By concentrating on distance and density based structures, this work restricts clusters to "natural" clusters (c.f. Duda et al., 2001) and therefore omits the axiom of richness where all partitions should be achievable. Thus, natural clusters consists of objects which are similar within clusters and dissimilar between clusters. "[Clusters] can be of arbitrary shapes [structures] and sizes in multidimensional pattern space. Each clustering criterion imposes a certain structure on the data, and if the data happen to conform to the requirements of a particular criterion, the true clusters are recovered." (Jain and Dubes, 1988). Here, the Databionic swarm (DBS) is used (Thrun, 2018) to find natural clusters without imposing a particular structure on the data contrary to conventional algorithms. An example of an algorithm imposing structures would be spectral clustering which searches for clusters with "chain-like or other intricate structures" (Duda et al., 2001) (see also Hennig et al., 2015). Spectral clustering lacks "robustness when there is little spatial separation between the clusters" (Handl et al., 2005). Such effects were made visible on simple artificial datasets for conventional algorithms (Thrun, 2018).

2 Distance-based Cluster Algorithm

The Databionic swarm (DBS) implements a swarm of agents interacting with one another and sensing their environment. DBS can adapt itself to structures of high-dimensional data such as natural clusters characterized by distance and density based structures in the data space (Thrun, 2018).

The DBS algorithm consists three modules. First, the projection method Pswarm, second the visualization technique of a topographic map based on the generalized U-matrix, and third, the clustering approach itself.

Pswarm is a swarm of intelligent agents called DataBots (Ultsch, 2000). It is a parameter-free focusing projection method of a polar swarm that exploits concepts of self-organization and swarm intelligence (Thrun, 2018). During construction of this type of projection, which is called learning phase and requires an annealing scheme, structure analysis shifts from global optimization to local distance preservation (focusing). Intelligent agents of Pswarm operate on

a toroid grid where positions are coded into polar coordinates allowing for a precise definition of their movement, neighborhood function and annealing scheme. The size of the grid and, in contrast to other focusing projection methods (e.g. Ultsch and Lötsch, 2017) the annealing scheme is data-driven and therefore, this method does not require any parameters. During learning, each DataBot moves across the grid or stay in its current position in the search for the most potent scent that means it searches for other agents carrying data with the most similar features to itself with a data-driven decreasing search radius (Thrun, 2018). Contrary to other projections methods and similar to the emergent self-organizing map, the Pswarm projection method does not possess a global objective function which allows the method to apply self-organization and swarm intelligence (Thrun, 2018).

Second, the projected points[2] are transformed to points on a discrete lattice; these points are called the best-matching units (BMUs) $bmu \in B \subset \mathbb{R}^2$ of the high-dimensional data points j. Then the generalized U*-matrix can be applied to the projected points by using a simplified emergent self-organizing map (ESOM) algorithm which is an unsupervised neural network (Thrun, 2018). The result is a topographic map with hypsometric tints (Thrun, Lerch, Lötsch, and Ultsch, 2016). Hypsometric tints are surface colors that represent ranges of elevation (see (Thrun et al., 2016)). Here, contour lines are combined with a specific color scale. The color scale is chosen to display various valleys, ridges, and basins: blue colors indicate small distances (sea level), green and brown colors indicate middle distances (low hills), and white colors indicate vast distances (high mountains covered with snow and ice). Valleys and basins represent clusters, and the watersheds of hills and mountains represent the borders between clusters. In this 3D landscape, the borders of the visualization are cyclically connected with a periodicity (L,C). A central problem in clustering is the correct estimation of the number of clusters. This is addressed by the topographic map which allows assessing the number of clusters (Thrun et al., 2016).

Third, in (Lötsch and Ultsch, 2014) it was shown that a single wall of the AU-matrix represents the actual distance information between two points in the high-dimensional space: the generalized U-matrix is the approximation of the abstract U-matrix (AU-matrix) (Lötsch and Ultsch, 2014). Voronoi cells around each projected point define the abstract U-matrix (AU-matrix) and generate a Delaunay graph. For every BMU all direct connections are weighted using the input-space distances $D(l, j)$, because on each border between two Voronoi cells a height is defined.

[2] Of DataBot positions on the hexagonal grid of Pswarm.

For the distances $D(l,j)$ the dynamic time warping (DTW) distances were calculated using the CRAN package in R "dtw" (Giorgino, 2009). "The DTW distance allows warping of the time axes to align the shapes of the two times series better. The two series can also be of different lengths. The optimal alignment is found by calculating the shortest warping path in the matrix of distances between all pairs of time points under several constraints. The point-wise distance is usually the Euclidean. The DTW is calculated using dynamic programming with time complexity O(n2)" (Mörchen, 2006).

Now, the distances between two points in the high-dimensional space is considered as the distance between two time series. All possible Delaunay paths $p_{j,l}$ between all points are calculated toroidal because the topographic map is toroidal. Then, the minimum of all possible path distances $p_{j,l}$ between a pair of points $\{j,l\} \in O$ in the output space is calculated as the shortest path $G(l,j,\mathcal{D})$ using the algorithm of Dijkstra resulting in a new high-dimensional distance $D^*(l,j)$. Here, the compact approach is used, where the two clusters with the minimal variance S are merged to together until given the number of clusters defined by the topographic map is reached.

Let $c_r \subset I$ and $c_q \subset I$ be two clusters such that $r,q \in \{1,\dots,k\}$ and $c_r \cap c_q = \{\}$ for $r \neq q$, and let the data points in the clusters be denoted by $j_i \in c_q$ and $l_i \in c_r$, with the cardinality of the sets being $k = |c_q|$ and $p = |c_r|$ and

$$\Delta Q(j,l) = \frac{k * p}{k + p} D^*(l,j)$$

then, the variance between two clusters is defined as

$$S(c_r, c_k) = \sum_{i=1,j=1,j\neq i}^{k,p} \Delta Q(j,l) \quad .$$

A dendrogram can be shown additionally. The clustering is valid if mountains do not partition clusters indicated by colored points of the same color and colored regions of points. The algorithm was run using the CRAN package in R "DatabionicSwarm".

3 Application: PPP-converted gross domestic product (GDP) per capita

The World GDP data set was logarithmized, and countries with missing values were not considered. As a result, 160 countries[3] remain for which the optimal alignment between two time series (Giorgino, 2009) is calculated and the DBS algorithm applied.

[3] For overview see (Leister, 2016, pp. 105-107)

In contrast to most conventional clustering algorithms, the topographic map allows identifying that clustering of the data is meaningless if the data contains no (natural) clusters (Thrun, 2018). Thus, Fig. 1 demonstrates a clear (natural) cluster structure. The homogeneity of the cluster structures of DBS is visualized in a silhouette plot in Fig. 3, and it is confirmed by the heat map (Fig. 2). In Fig. 4 the Classification and Regression Tree (CART) analysis is shown. The clusters are defined mainly by an event that occurred in 2001. The rules generated from the CART are presented in Table 1, and applied as colored labels to a world map in Fig. 5 with the same colored points as Fig. 1.

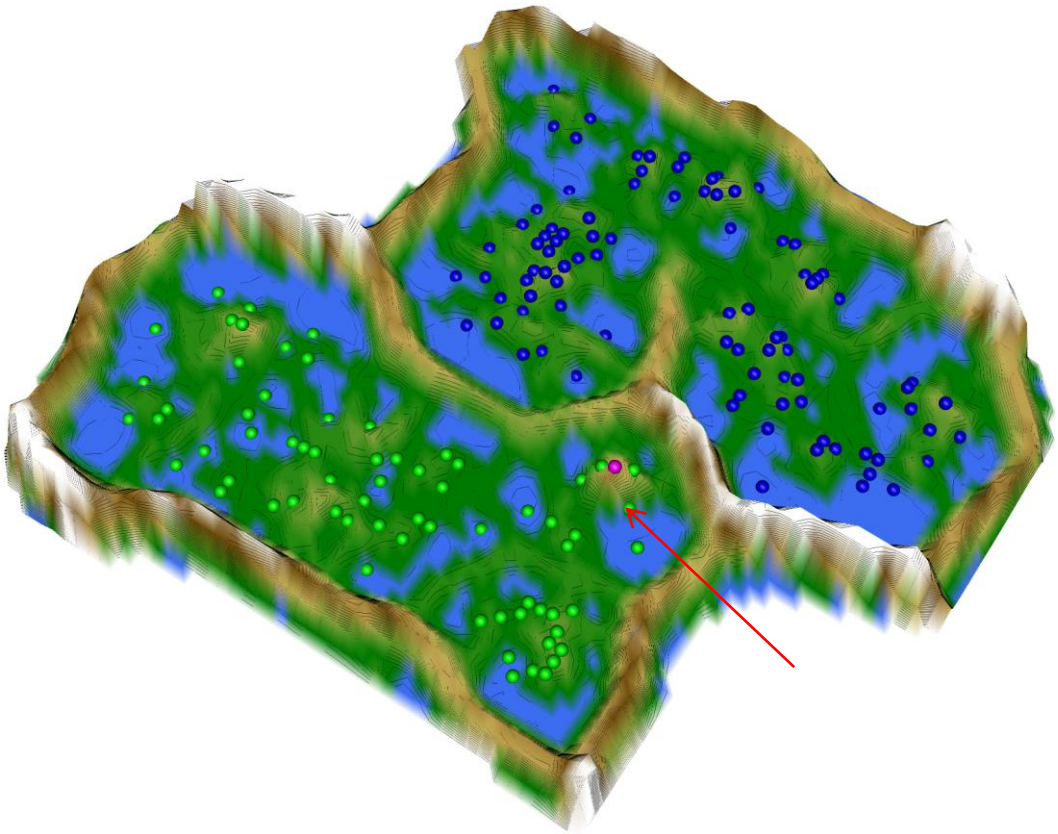

Fig. 1. Topographic map of the DBS clustering of the World GDP data set shows two distinctive clusters. There is one outlier, colored in magenta and marked with a red arrow. The visualization was generated by the CRAN package in R "DatabionicSwarm".

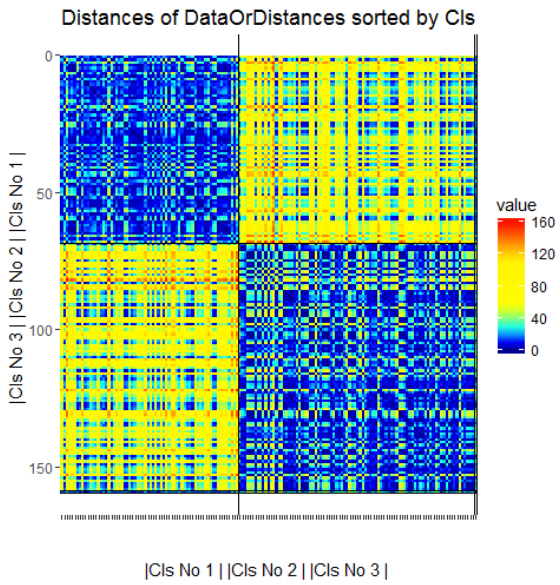

Fig. 2. Heatmap of the dynamic time warping (DTW) distances for the World GDP data set shows a small variance of intracluster distance. The visualization was generated by the CRAN package in R "DataVisualizations".

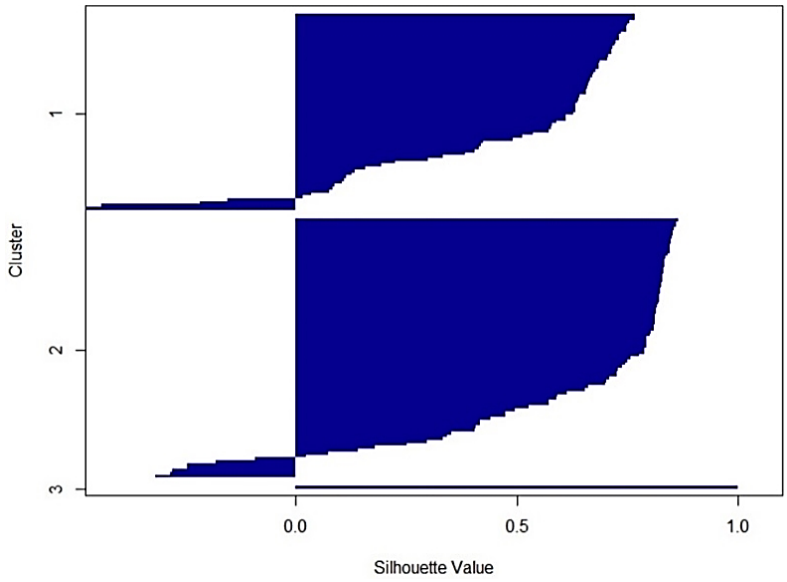

Fig. 3. Silhouette plot of the DBS clustering results for the World GDP data set indicates that data points (y-axis) above a value of 0.5 (x-axis) have been assigned to an appropriate cluster. The visualization was generated by the CRAN package in R "DataVisualizations".

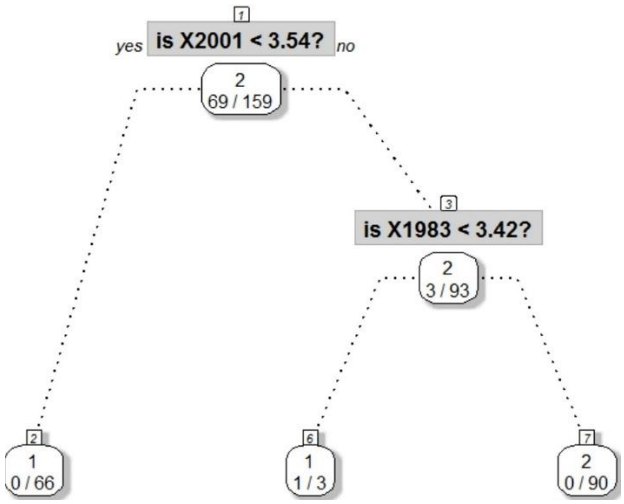

Fig. 4. Classification and Regression Tree (CART) analysis rules for the clusters. The two main clusters are defined only by an event in 2001.

Table 1. The CART rules based on Fig. 4 in which cluster of Fig.1 is used. Egypt, Micronesia and the outlier Equatorial Guinea classified incorrectly by these two rules.

Rule No.	DBS Cluster No.	No. of Nations	Rules
R1	1	66	BIP lower than 3469 U in the year 2001
R2	2	93	BIP higher than 3469 U in the year 2001

Notes: Abbreviations - U: PPP-converted GDP per capita.

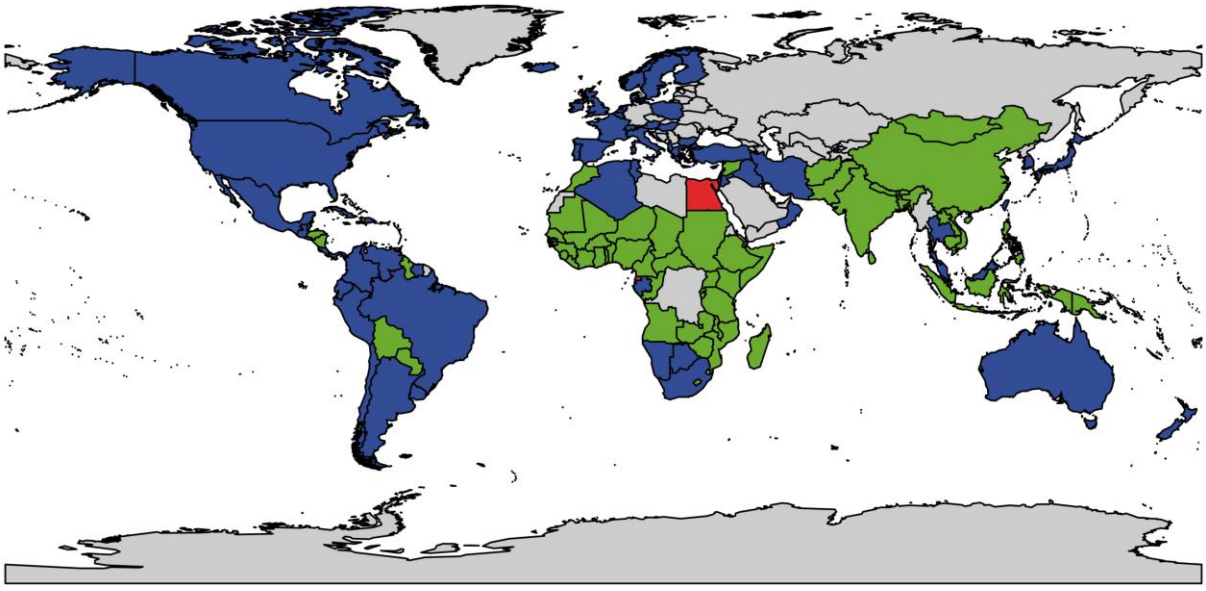

Fig. 5. Two rules of Table 1 classify countries in blue and green. For grey countries, no data was available in (Leister, 2016). The rules result from the clustering of Fig. 1. In red are the Outlier Equatorial Guinea as well as the incorrectly classified countries Egypt and Micronesia by these two rules.

4 Discussion

Regional cluster analysis on GDP datasets was performed for Latin American countries in (Redelico et al., 2009) and European countries in (Gallo and Ertur, 2003). To the knowledge of the author, no cluster analysis of the whole world was performed with the goal to explain the clusters by rules and through a spatial world map (Fig. 5). Here, both clusters found by the Databionic swarm are spatially separated. The first cluster consists mostly of African and Asian countries and the second clusters of industrialized countries of predominantly Europe and America. Due to the correlations between the human development index (HDI) and PPP-converted log(GDP) per capita shown in Fig. 3.2 on page 69 in (UNDP, 2003), the second cluster in Fig. 5 is highly similar to the HDI map of Fig. 1 in (Birdsall and Birdsall, 2005) with HDI higher than 0.7.

The two rules of the (CART) analysis which are presented in Table 1, demonstrate that the clusters are defined by an event that occurred in 2001, which could be the crashing of airplanes into the World Trade Center. In its aftermath, "the world economy was experiencing its first synchronized global recession in a quarter-century" (Makinen, 2002, p. 17). Therefore, the results indicate that the first cluster of African and Asian countries was unaffected by this event, and the second cluster of American and European countries was affected. As published in Vollmer et al. (2013), the GDP is sensitive by economic shocks, e.g., oil-price of excludingly oil-exporting countries or countries with a low number of inhabitants (Vollmer et al., 2013). The data regarding the PPP-converted GDP per capita of Egypt may be misrepresented, because "during the twentieth century the population of Egypt has increased by more than 5 times" (El Araby, 2002). The outlier in Fig. 1 describes the data of Equatorial Guinea. This small country with an area of 28,000 square kilometers is mostly based on oil and is, one of sub-Saharan Africa's largest oil producers. The Federated States of Micronesia is a subregion of Oceania has only a low number of inhabitants (105.000). Thus, it could also be an outlier.

Conclusions

The Databionic swarm (DBS) resulted in a coherent spatiotemporal clustering of the multivariate time series of the PPP-converted gross domestic product (GDP) per capita of 160 countries in the years 1970 to 2010. It seems that 157 countries can be classified by using two rules extracted from the CART with only one threshold for the GDP in the year 2001. This indicates that the economic achievement of these countries were profoundly affected in the year 2001. DBS can be by downloaded as the CRAN package in R "DatabionicSwarm".

References

Birdsall, S. & Birdsall, W. (2005). *Geography matters: Mapping human development and digital access. 2005*. doi:10.5210/fm.v10i10.1281.

Duda, R. O., Hart, P. E. & Stork, D. G. (2001). *Pattern Classification (2nd ed.)*. Ney York, USA: John Wiley & Sons.

El Araby, M. (2002). Urban growth and environmental degradation: The case of Cairo, Egypt. *Cities, 19*(6), 389-400.

Gallo, J. & Ertur, C. (2003). Exploratory spatial data analysis of the distribution of regional per capita GDP in Europe, 1980–1995. *Papers in regional science, 82*(2), 175-201.

Giorgino, T. (2009). Computing and visualizing dynamic time warping alignments in R: the dtw package. *Journal of statistical Software, 31*(7), 1-24.

Handl, J., Knowles, J. & Kell, D. B. (2005). Computational cluster validation in post-genomic data analysis. *Bioinformatics, 21*(15), 3201-3212.

Hennig, C., Meila, M., Murtagh, F. & Rocci, R. (2015). *Handbook of cluster analysis*. New York, USA: CRC Press.

Heston, A., Summers, R. & Aten, B. (2012). Penn World Table Version 7.1 Center for International Comparisons of Production. *Income and Prices at the University of Pennsylvania*.

Jain, A. K. & Dubes, R. C. (1988). *Algorithms for Clustering Data* (Vol. 3). Englewood Cliffs, New Jersey, USA: Prentice Hall College Div.

Kleinberg, J. (2003). *An impossibility theorem for clustering.* Paper presented at the Advances in neural information processing systems.

Leister, A. M. (2016). *Hidden Markov models: Estimation theory and economic applications.* Dissertation, Marburg, Germany.

Lötsch, J. & Ultsch, A. (2014). *Exploiting the Structures of the U-Matrix.* Paper presented at the Advances in Self-Organizing Maps and Learning Vector Quantization, Mittweida, Germany.

Makinen, G. (2002). *The economic effects of 9/11: A retrospective assessment*. Library of congress, Washington D.C. congressional research service.

Mazumdar, K. (2000). Causal flow between human well-being and per capita real gross domestic product. *Social Indicators Research, 50*(3), 297-313.

Mörchen, F. (2006). *Time series knowledge mining*. Dissertation, Marburg, Germany: Citeseer/Görich & Weiershäuser.

Redelico, F. O., Proto, A. N. & Ausloos, M. (2009). Hierarchical structures in the Gross Domestic Product per capita fluctuation in Latin American countries. *Physica A: Statistical Mechanics and its Applications, 388*(17), 3527-3535.

Rogoff, K. (1996). The purchasing power parity puzzle. *Journal of Economic literature, 34*(2), 647-668.

Thrun, M. C. (2018). *Projection Based Clustering through Self-Organization and Swarm Intelligence.* A. Ultsch & E. Hüllermeier Adv.. Dissertation, Heidelberg: Springer.

Thrun, M. C., Lerch, F., Lötsch, J., & Ultsch, A. (2016). *Visualization and 3D Printing of Multivariate Data of Biomarkers*. Paper presented at the International Conference on Computer Graphics, Visualization and Computer Vision (WSCG), Plzen.

Ultsch, A. (2000). *Clustering with DataBots*. Paper presented at the Int. Conf. Advances in Intelligent Systems Theory and Applications (AISTA), Canberra, Australia.

Ultsch, A. & Lötsch, J. (2017). Machine-learned cluster identification in high-dimensional data. *Journal of biomedical informatics, 66*, 95–104. DOI:10.1016/j.jbi.2016.12.011.

UNDP. (2003). *Human development Report: Millennium Development Goals: a compact among nations to end poverty*. United Nations Development Programme (UNDP), New York: Oxford University Press.

Vollmer, S., Holzmann, H. & Schwaiger, F. (2013). Peaks vs components. *Review of Development Economics, 17*(2), 352-364.