EBER
Entrepreneurial Business and Economics Review

2019, Vol. 7, No. 1

OPEN ACCESS

10.15678/EBER.2019.070113

Cluster Analysis of Per Capita Gross Domestic Products[1]

Michael C. Thrun

ABSTRACT

Objective: The purpose of this article is to show the value of exploratory data analysis performed on the multivariate time series dataset of gross domestic products per capita (GDP) of 160 countries for the years 1970-2010. New knowledge can be derived by applying cluster analysis to the time series of GDP to show how patterns in GDP can be explained in a data-driven way.

Research Design & Methods: Patterns characterised by distance and density based structures were found in a topographic map by using dynamic time warping distances with the Databionic swarm (DBS) [1]. The topographic map represents a 3D landscape of data structures. Looking at the topographic map, the number of clusters was derived. Then, a DBS clustering was performed and the quality of the clustering was verified.

Findings: Two clusters are identified in the topographic map. The rules deduced from classification and regression tree (CART) show that the clusters are defined by an event occurring in 2001 at which time the world economy was experiencing its first synchronised global recession in a quarter-century. Geographically, the first cluster mostly of African and Asian countries and the second cluster consists mostly of European and American countries.

Implications & Recommendations: DBS can be used even by non-professionals in the field of data mining and knowledge discovery. DBS is the first swarm-based clustering technique that shows emergent properties while exploiting concepts of swarm intelligence, self-organisation, and game theory.

Contribution & Value Added: To the knowledge of the author it is the first time that worldwide similarities between 160 countries in GDP time series for the years 1970-2010 have been investigated in a topical context.

Article type:	research article
Keywords:	machine learning; cluster analysis; swarm intelligence; visualisation, self-organisation; gross domestic product
JEL codes:	O47, F01, C380

Received: 13 May 2018 Revised: 2 January 2019 Accepted: 7 January 2019

Suggested citation:

Thrun, M.C. (2019). Cluster Analysis of Per Capita Gross Domestic Products. *Entrepreneurial Business and Economics Review,* 7(1), 217-231. https://doi.org/10.15678/EBER.2019.070112

[1] It should be noted that on page 129, (Thrun, 2018) the dataset was used as one out of twenty examples to indicate that DBS is able to find structures in a variety of cases.

INTRODUCTION

The multivariate time series inspected in this work covers repeated measures of the gross domestic product (GDP) of 190 countries published in Heston, Summers and Aten (2012) but not every time series could be used for cluster analysis. 'GDP measures the monetary value of final goods and services' (Callen, 2008). Final goods are all commodities currently produced, exchanged and consumed although this definition is controversial (England, 1998). Thus, GDP is an indicator of the economic performance of a country (Mazumdar, 2000). Each country's data has to be converted into a common currency to make international comparisons. 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 the database developed by Heston *et al*. (2012) by selecting the PPP-converted GDP per capita for the years from 1970 to 2010 by Leister (2016). With the help of exploratory data science, this work shows how to search for meaningful structures in GDP. The World-GDP data set will be investigated in the context of economic similarity between nations by combining dimensionality reduction with cluster analysis. New and valid knowledge will be extracted from the structures defined by a hybrid algorithm consisting of an artificial swarm and a self-organising map.

In market segmentation, cluster analysis was applied to countries in which some indicators were defined by percentages of GDP (Day, Fox, & Huszagh, 1988; Kantar, Deviren, & Keskin, 2014; Liapis, Rovolis, Galanos, & Thalassinos, 2013; Powell & Barrientos, 2004) or correlations between GDP and various variables or cross-correlations of already clustered countries were investigated (Ausloos & Lambiotte, 2007; Franceschini, Galetto, Maisano, & Mastrogiacomo, 2010; Furnham, Kirkcaldy, & Lynn, 1996). Alternative approaches with the goal to cluster countries were performed using GDP and other variables at the same point in time (Michinaka, Tachibana, & Turner, 2011). To the knowledge of the author, neither a visualisation of structures based on GDP using dimensionality reduction was performed, nor was a data-driven approach used to explain such structures.

The methods used in cluster analysis 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, 2015, p. 705). Additionally, Kleinberg showed for a set of three simple axioms called 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, Hart, & Stork, 2001, p. 539) and therefore omits the axiom of richness where all partitions should be achievable. Thus, natural clusters consist 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 recovere' (Jain & Dubes, 1988, p. 91). Here, the Databionic swarm (DBS) is used to find natural clusters without imposing a particular structure on the data contrary to conventional algorithms (Thrun, 2018). The purpose of this work is to show that the cluster structures found with DBS are meaningful, new and interesting, whereas in

Thrun (2018) the cluster structures were investigated from a methodological point of view, e.g. described in (Behnisch & Ultsch, 2015). 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, p. 582) (see also Hennig, 2015, p. 10). Spectral clustering lacks 'robustness when there is little spatial separation between the clusters' (Handl, Knowles, & Kell, 2005, p. 3202). For conventional clustering algorithms, such effects were made visible on simple artificial datasets (Thrun, 2018, pp. 118-124).

This work is structured as follows. In the next section, the distance-based visualisation and clustering algorithm of the Databionic swam (DBS) is explained. In the third section, DBS is applied to the world GDP dataset. The results are discussed in the fourth section leading the conclusion in the last section.

MATERIAL AND METHODS

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 characterised by distance and density based structures in the data space (Thrun, 2018). The algorithm consists of three modules: the non-linear projection method Pswarm, the visualisation technique of a topographic map based on the generalised U-matrix and 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-organisation and swarm intelligence (Thrun, 2018). During construction of this type of projection, which is called the learning phase and requires an annealing scheme, structure analysis shifts from global optimisation 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, neighbourhood function and annealing scheme. The size of the grid and, in contrast to other focusing projection methods (e.g. Demartines & Hérault, 1995; Ultsch & Lötsch, 2017; Van der Maaten & Hinton, 2008), the annealing scheme are data-driven, and therefore, this method does not require any parameters. During learning, each DataBot moves across the grid or stays 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). The movement of every DataBot is modelled using an approach of game theory, and the radius decreases only if a Nash equilibrium is found (Nash, 1951). Contrary to other projection methods and similar to the emergent self-organising map, the Pswarm projection method does not possess a global objective function which allows the method to apply self-organisation and swarm intelligence (Thrun, 2018).

In the second module, the projected points $\{l, j\}$ 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 $\{l, j\}$. Then the generalised U matrix can be applied to the projected points by using a simplified emergent self-organising map (ESOM) algorithm which is an unsupervised neural network (Thrun, 2018). The result is a topographic map with hypsometric tints (Thrun, Lerch, Lötsch, & Ultsch, 2016). Hypsometric tints are surface colours that represent ranges of elevation (see Thrun *et al.*, 2016). Here, contour lines are combined with a specific colour scale. The colour scale is chosen to display various valleys, ridges, and basins: blue colours indicate small distances (sea level), green and brown colours indicate middle distances (low hills), and shades of white colours 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 visualisation 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 to assess the number of clusters (Thrun *et al.*, 2016).

The third module is the clustering approach itself. In (Lötsch & Ultsch, 2014) it was shown that a single wall of the abstract U-matrix (AU-matrix) represents the actual distance $D(l,j)$ information between two points in the high-dimensional space: the generalised U-matrix is the approximation of the AU-matrix (Lötsch & Ultsch, 2014). Voronoi cells around each projected point define the abstract U-matrix (AU-matrix) and generate a Delaunay graph $\mathcal{D}$. 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.

For the distance $D(l,j)$ 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 one. The DTW is calculated using dynamic programming with time complexity $O(n^2)$' (Mörchen, 2006, p. 24).

Now, the distances between two points in the high-dimensional space are considered as the distance between two time series. All possible weighted Delaunay paths $p_{l,j}$ between all points are calculated toroidal because the topographic map is toroidal. Then, the minimum of all possible path distances between a pair of points $\{l,j\} \in O$ in the output space O is calculated as the shortest path $G(l,j,\mathcal{D})$ using the algorithm of (Dijkstra, 1959) 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 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

$$\Delta Q(j,l) = \frac{k * p}{k + p} D^*(l,j) \tag{1}$$

where:

(l,j) - the data points in the clusters be denoted by $j_i \in c_q$ and $l_i \in c_r$;

k - the cardinality $|c_q|$ of the first set;

p - the cardinality $|c_r|$ of the second set;

D^* - high-dimensional distance based on weighted shortest paths in the Delaunay Graph.

then, the variance S between two clusters is defined as

$$S(c_r, c_k) = \sum_{i=1,j=1,j\neq i}^{k,p} \Delta Q\,(l,j) \tag{2}$$

A dendrogram can be shown additionally. The clustering is valid if mountains do not partition clusters indicated by coloured points of the same colour and coloured regions of points. The algorithm was run using the CRAN package in R ‘Databion-icSwarm’.

RESULTS

The World GDP data set of Leister (2016, pp. 105-107) was logarithmised, and countries with missing values were not considered. As a result, 160 time series of countries remain for which the optimal alignment between every two two time series is calculated using the R package ‘dtw’ on CRAN (Giorgino, 2009). The cluster analysis approach of the Databionic swarm (DBS) has one parameter defining either a clustering for distance or density based cluster structures. Thus, the probability density distribution of the dynamic time warping (DTW) distances is investigated using the Pareto density estimation (PDE) which is particularly suitable for finding groups in data (Ultsch, 2005). A Gaussian mixture model (GMM) can be calculated (Figure 1). There, the first mode consists of smaller distances and the second mode of larger distances. Both modes are drawn in blue, their superposition in red and the probability density function estimated by PDE in black.

Comparing the GMM to the percentiles of the distances, the Quantile-Quantile plot shows a good fit, meaning that the distances can be clearly separated in larger inter-cluster distances and smaller intra-cluster distances. Therefore, the dataset has a clear distance structure. With this information, the clustering of the DBS algorithm is computed.

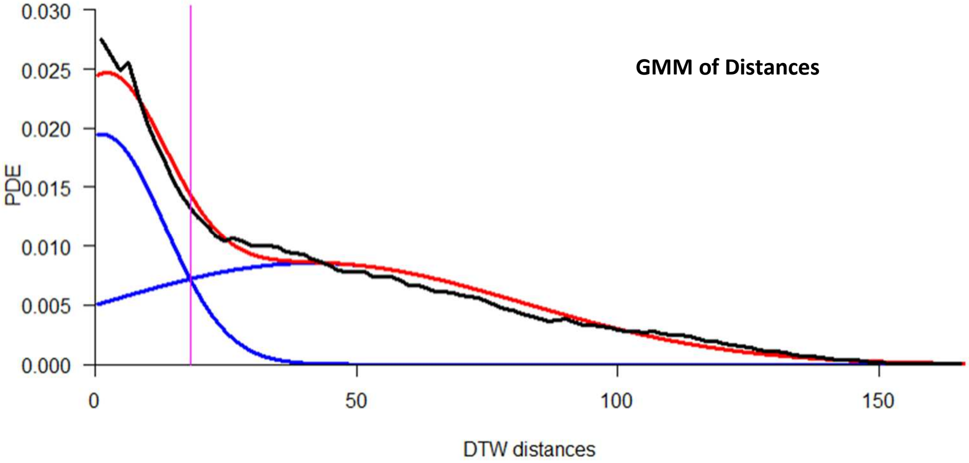

Figure 1a. GMM (red) is based on the pdf of DTW distances (black) between countries estimated by PDE

Source: The visualisation was generated using the R package ‘AdaptGauss’ available on CRAN (Ultsch, Thrun, Hansen-Goos, & Lötsch, 2015).

In contrast to most conventional clustering algorithms, the topographic map allows to visualise high-dimensional distances and densities between the projected points identifying that clustering of the data is meaningless if no structures are visible (Thrun, 2018). In this 3D landscape the heights and colour scale are chosen in such a way that small heights indicate small distances (sea level), middle heights indicate middle distances (low hills), and large heights indicate vast distances. Valleys and basins represent clusters, and the watersheds of hills and mountains represent the borders between clusters. Thus, Figure 2 demonstrates a clear (natural) cluster structure. Additionally, the quality of the clustering of DBS is confirmed by the heat map (Figure 3) which shows small intra-cluster distances in every cluster and high inter-cluster distances between the two clusters. The Silhouette plot in Figure 4 indicates a good spherical cluster structure for values above 0.5. This corroborates the results illustrated in Figure 2 showing that countries in the same cluster are similar to each other and countries being in different clusters are not. Contrary to the non-linear approach of the projection method Pswarm, a linear projection method into two dimensions called independent component analysis (ICA) (Hyvärinen, Karhunen, & Oja, 2004) is unable to capture the cluster structure (Figure 5). This indicates that a linear model would be unable to distinguish the distance-based structures of the World GDP dataset.

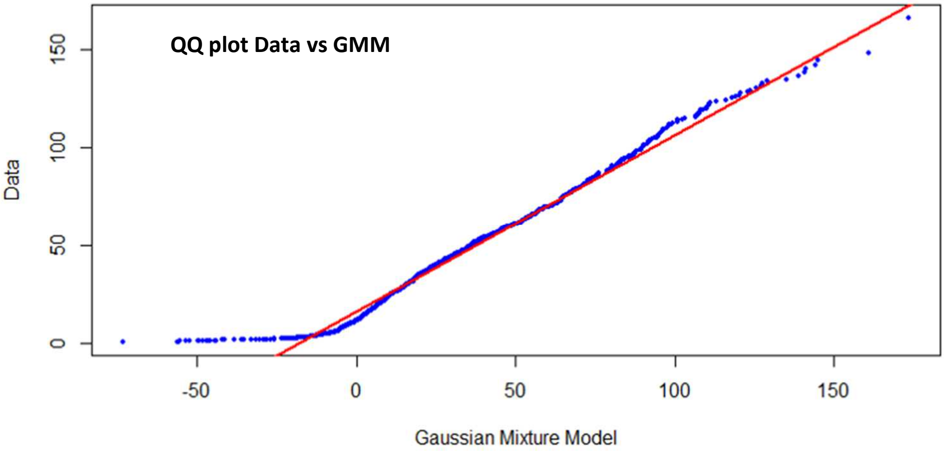

Figure 1b. The QQ plot shows a good match between the data of distance and the GMM through the straight line

Source: The visualisation was generated using the R package 'AdaptGaus' available on CRAN (Ultsch et al., 2015).

In Figure 6 the result of the Classification and Regression Tree (CART) algorithm is presented. The clusters are defined mainly by an event that occurred in 2001 if one follows the path from the root to leaf in the tree. The rules generated from the CART are presented in Table 1, and applied as coloured labels to the world map in Figure 7 with the same coloured points as in Figure 2. By using the CART classification, the two main classes have different distributions of GDP which is visualized with the Mirrored Density plot or so-called MD-plot (Thrun & Utsch, 2019). The MD-plot is depicted in Figure 8.

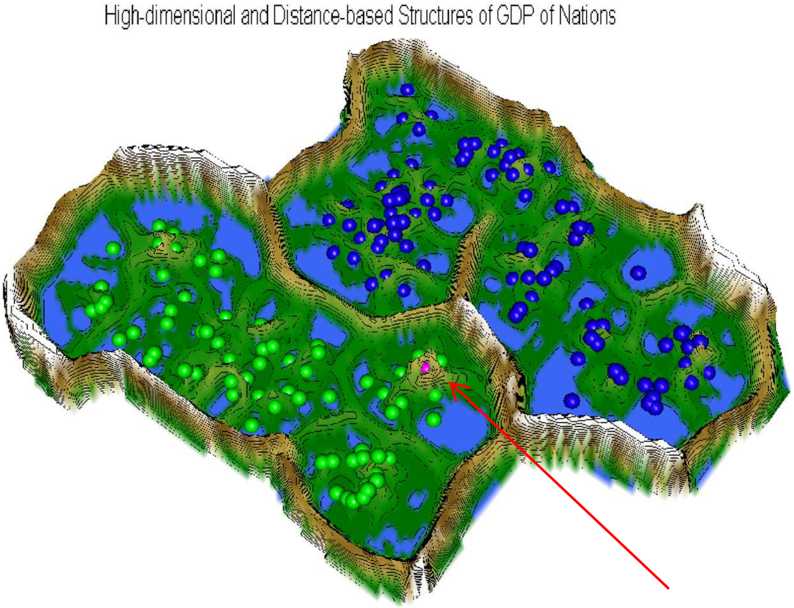
High-dimensional and Distance-based Structures of GDP of Nations

Figure 2. The topographic map of the DBS clustering of the World GDP data set shows two distinctive clusters, c.f. (Thrun, 2018). There is one outlier, coloured in magenta and marked with a red arrow

Source: The visualisation was generated using the R package 'DatabionicSwarm' available on CRAN (Thrun, 2018).

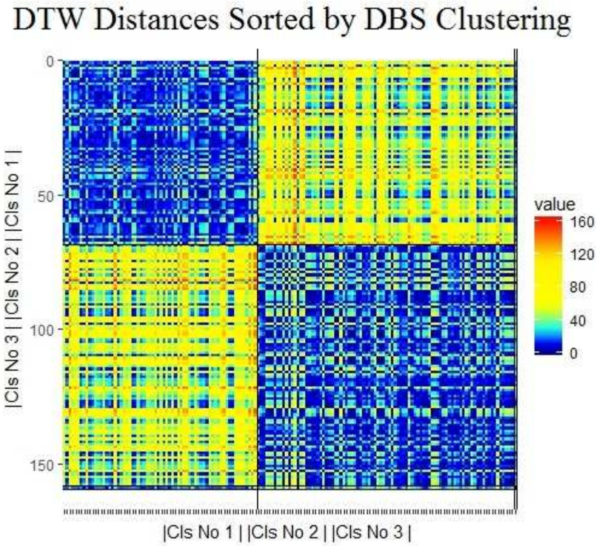

Figure 3. The heatmap of the DTW distances for the World GDP dataset, c.f. (Thrun, 2018), shows a small variance of intracluster distance in blue colours and large inter-cluster distances in yellow and red colours

Source: The visualisation was generated using the R package 'DataVisualizations' available on CRAN (Thrun & Ultsch, 2018).

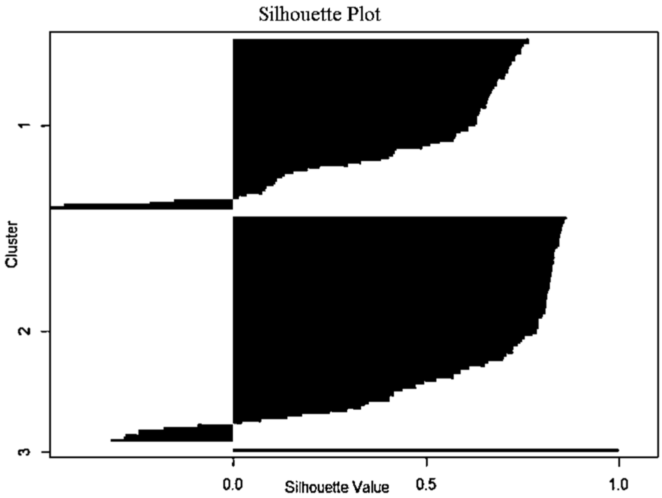

Figure 4. The silhouette plot of the DBS clustering results for the World GDP data set indicates that data points (y-axis) above a value of 0.5 (x-axis) have been assigned to an appropriate cluster, c.f. (Thrun, 2018)

Source: The visualisation was generated using the R package 'DataVisualizations' available on CRAN (Thrun & Ultsch, 2018).

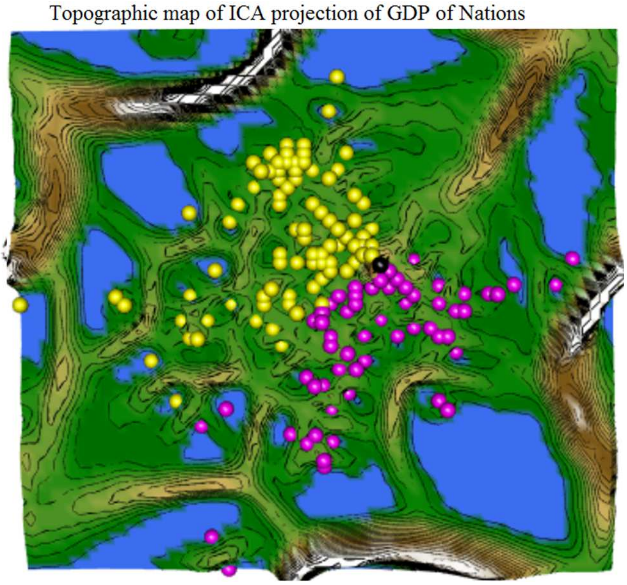

Figure 5. The linear projection of the independent component analysis (ICA) is unable to distinguish the clusters, even if generalised U-matrix is applied to generate a topographic map out of the two-dimensional projection

Source: The visualisation was generated using the R package 'ProjectionBasedClustering' (Thrun & Ultsch, 2017) and 'GeneralizedUmatrix' (Ultsch & Thrun, 2017) available on CRAN.

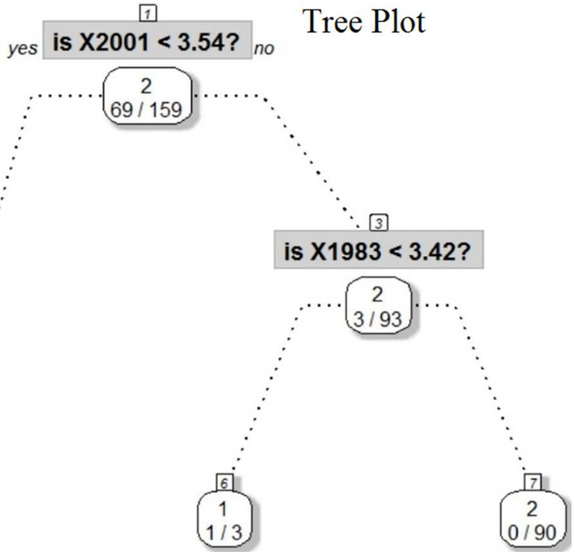

Figure 6. Classification and Regression Tree (Cart) analysis reveal rules for the clusters, c.f. (Thrun, 2018). The two main clusters are defined only by an event in 2001

Source: The visualisation was generated using the R package 'rpart' available on CRAN (Therneau, Atkinson, Ripley, & Ripley, 2018).

Table 1. The CART Rules Based on Figure 4 in Which Cluster of Figure 1 is Used (*)

DBS Cluster No./Rule No.	Clusterwise Median Distance	Medoids	No. of Nations	Rules
1/R1	12.2	Sudan	66	GDP lower than 3469 Y in the year 2001
2/R2	12.6	Taiwan	93	GDP higher than 3469 Y in the year 2001

* Egypt, Micronesia and the outlier Equatorial Guinea classified incorrectly by these two rules. Abbreviations – Y: PPP-converted GDP per capita.

Source: own elaboration using the R programming language [R Development Core Team, 2008].

DISCUSSION

Regional cluster analysis on GDP datasets was performed for Latin American countries in Redelico, Proto and Ausloos (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 (Figure 7). Here, both clusters found by the Databionic swarm are spatially separated.

The distribution analysis of distances shows two Gaussian modes. The analysis indicates that by choosing the DTW distance measure, high-dimensional distance-based structures can be found because the first mode indicates small intra-cluster distances and the second mode large inter-cluster distances. The values of the cluster-wise median distance support this indication (Table 1). The DBS is able to visualise these structures using dimensionality reduction and able to generate a clustering based on the DTW distances.

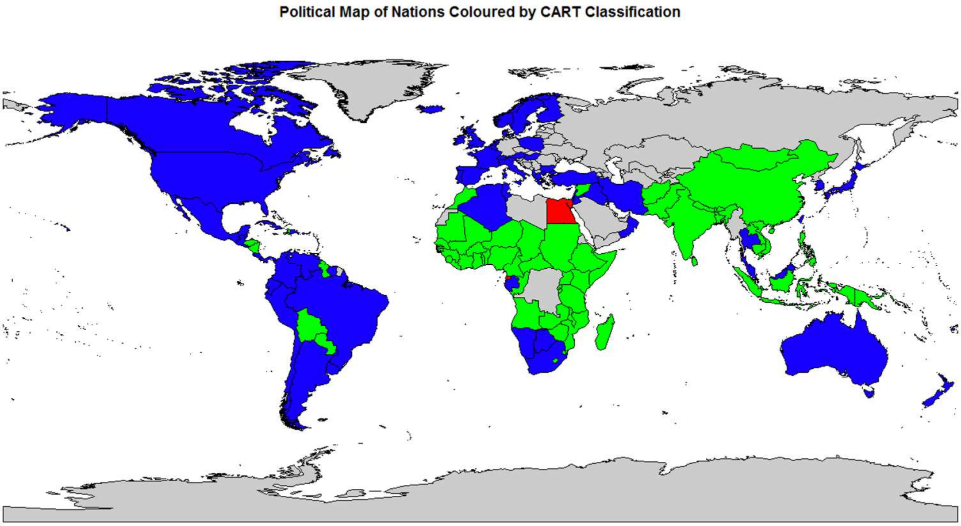

Figure 7. Two rules of Table 1 classify countries in a political map of blue and green countries. For grey countries, no data was available in Leister (2016), e.g. Balkan countries. The rules result from the clustering of Figure 1. In red there are the Outlier Equatorial Guinea as well as the incorrectly classified countries Egypt and Micronesia

Source: The visualisation was generated using the R package 'DataVisualizations' available on CRAN (Thrun & Ultsch, 2018).

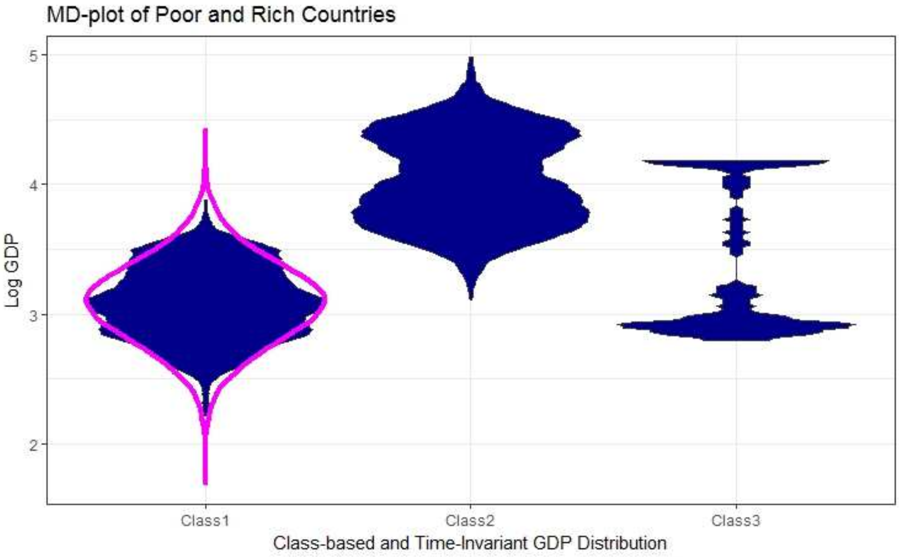

Figure 8. PDE of class-based probability density distribution of log-transformed GDP for all years combined shows that the classes can be roughly distinct in poor and rich countries

Source: The Mirrored Density plot (MD-plot) was generated using the R package 'DataVisualizations' available on CRAN (Thrun & Ultsch, 2018).

The extracted cluster structures by DBS are meaningful: The first cluster consists mostly of African and Asian countries and the second cluster of industrialised countries predominantly in Europe and America. Due to the correlations between the human development index (HDI) and PPP-converted log(GDP) per capita shown in Figure 3.2 on page 69 in (UNDP, 2003, p. 69, Fig. 3.2), the second cluster in Figure 7 is highly similar to the HDI map of Figure 1 in Birdsall and Birdsall (2005, Figure 1) with HDI higher than 0.7.

The analysis was performed exploratively, meaning that contextual information was disregarded and GDP was clustered with a data-driven approach. It is surprising that neither hunger periods (e.g. Somalia, Ethiopia, Nigeria) nor war periods (Iran, Iraq) seem to affect the analysis because more detailed structures did not exist and additional outliers could not be found. In future research, the typical path of a time series of GDP would be an interesting point to investigate in order to understand why such events do not affect GDP strongly enough to be seen in the cluster structures. With the current analysis, homogeneity of clusters indicates that either a clustering analysis using DTW distances is not sensitive enough to be influenced by war or hunger or GDP itself is not affected strongly enough by such events.

If one follows every path from the root to the leaves of the tree, the resulting 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 aeroplanes into the World Trade Center. In aftermath of that event, ‘the world economy was experiencing its first synchronised 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, Holzmann and Schwaiger (2013), GDP can be sensitive to economic shocks, e.g., oil-price of exclusively oil-exporting countries or countries with a low number of inhabitants (Vollmer, Holzmann, & Schwaiger, 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 Figure 1 describes the data of Equatorial Guinea. This small country with an area of 28 000 square kilometres 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 and has only a low number of inhabitants (105 000). Thus, it could also be an outlier.

The choice of distance measure was strictly based on data-driven assumptions (Figure 1 and Table 1) resulting in meaningful structures. The DTW distance copes with time deformations and different speeds associated with time series (Müller, 2007) but can enforce delays in reaction to shocks. Thus, DTW does not necessarily account for smaller shocks at the same time. The CART tree indicates that the shock in 2001 was massive, resulting in the most prominent property defining the clustering. However, keeping the ugly-duckling theorem in mind (Watanabe, 1969, pp. 376-379), clustering is always biased towards the choice of properties. In exploratory data science it is preferable to make such a choice based on data but if a specific hypothesis would be pursued another choice of distance measure could result in other insights about the data.

CONCLUSIONS

This work shows the merits of applying exploratory data analysis on data before pursuing a specific hypothesis. The clustering derived from the Databionic swarm (DBS) resulted in coherent spatiotemporal clustering of the multivariate time series of the PPP-converted gross domestic products (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 CART with only one threshold for GDP in the year 2001. This indicates that the economic perfomance of these countries were profoundly affected in the year 2001. The knowledge of the existence of meaningful structures in GDP is vital in the pursuance of a specific hypothesis because it should be tested on the two main clusters separately if GDP takes part in an analysis. As a side-effect, a data-driven approach for defining poor and rich countries by log GDP distributions was defined. This approach is clearly non-linear and could not be applied without searching for structures beforehand. DBS can be downloaded as the R package ‘DatabionicSwarm’ on CRAN.

REFERENCES

Ausloos, M., & Lambiotte, R. (2007). Clusters or networks of economies? A macroeconomy study through gross domestic product. *Physica A: Statistical Mechanics and its Applications, 382*(1), 16-21. https://doi.org/10.1016/j.physa.2007.02.005

Behnisch, M., & Ultsch, A. (2015). Knowledge Discovery in Spatial Planning Data: A Concept for Cluster Understanding *Computational Approaches for Urban Environments* (pp. 49-75). Springer.

Birdsall, S., & Birdsall, W. (2005). Geography matters: Mapping human development and digital access. *First Monday, 10*(10). https://doi.org/10.5210/fm.v10i10.1281

Callen, T. (2008). What Is Gross Domestic Product?. *Finance & Development, 45*(4), 48-49.

Day, E., Fox, R.J., & Huszagh, S.M. (1988). Segmenting the global market for industrial goods: issues and implications. *International Marketing Review, 5*(3), 14-27.

Demartines, P., & Hérault, J. (1995). *CCA:"Curvilinear component analysis".* Paper presented at the 15° Colloque sur le traitement du signal et des images, France 18-21 September..

Dijkstra, E.W. (1959). A note on two problems in connexion with graphs. *Numerische Mathematik, 1*(1), 269-271.

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

El Araby, M. (2002). Urban growth and environmental degradation: The case of Cairo, Egypt. *Cities, 19*(6), 389-400. https://doi.org/10.1016/S0264-2751(02)00069-0

England, R.W. (1998). Measurement of social well-being: alternatives to gross domestic product. *Ecological Economics, 25*(1), 89-103.

Franceschini, F., Galetto, M., Maisano, D., & Mastrogiacomo, L. (2010). Clustering of European countries based on ISO 9000 certification diffusion. *International Journal of Quality & Reliability Management, 27*(5), 558-575. https://doi.org/10.1108/02656711011043535

Furnham, A., Kirkcaldy, B.D., & Lynn, R. (1996). Attitudinal correlates of national wealth. *Personality and Individual Differences, 21*(3), 345-353.

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. https://doi.org/10.1111/j.1435-5597.2003.tb00010.x

Giorgino, T. (2009). Computing and visualizing dynamic time warping alignments in R: the dtw package. *Journal of Statistical Software, 31*(7), 1-24. https://doi.org/10.18637/jss.v031.i07

Handl, J., Knowles, J., & Kell, D.B. (2005). Computational cluster validation in post-genomic data analysis. *Bioinformatics, 21*(15), 3201-3212. https://doi.org/10.1093/bioinformatics/bti517

Hennig, C., *et al*. (Hg.) (2015). *Handbook of cluster analysis.* New York, USA: Chapman & Hall/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*.

Hyvärinen, A., Karhunen, J., & Oja, E. (2004). *Independent component analysis* (Vol. 46).

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

Kantar, E., Deviren, B., & Keskin, M. (2014). Hierarchical structure of the European countries based on debts as a percentage of GDP during the 2000–2011 period. *Physica A: Statistical Mechanics and its Applications, 414*, 95-107.

Kleinberg, J. (2003). *An impossibility theorem for clustering.* Paper presented at the Advances in neural information processing systems, (Vol. 15, pp. 463-470). MIT Press, Vancouver, British Columbia, Canada December 9-14.

Leister, A.M. (2016). *Hidden Markov models: Estimation theory and economic applications.* Marburg: Philipps-Universität Marburg.

Liapis, K., Rovolis, A., Galanos, C., & Thalassinos, E. (2013). The Clusters of Economic Similarities between EU Countries: A View Under Recent Financial and Debt Crisis. *European Research Studies, 16*(1), 41.

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, July 2–4.

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

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

Michinaka, T., Tachibana, S., & Turner, J.A. (2011). Estimating price and income elasticities of demand for forest products: cluster analysis used as a tool in grouping. *Forest Policy and Economics, 13*(6), 435-445.

Mörchen, F. (2006). *Time series knowledge mining*. Marburg, Germany: Philipps-Universität Marburg, Görich & Weiershäuser.

Müller, M. (2007). *Information Retrieval for Music and Motion, Chapter Dynamic Time Warping*. Heidelberg, Germany: Springer.

Nash, J.F. (1951). Non-cooperative games. *Annals of Mathematics*, 286-295.

Powell, M., & Barrientos, A. (2004). Welfare regimes and the welfare mix. *European Journal of Political Research, 43*(1), 83-105. https://doi.org/10.1111/j.1475-6765.2004.00146.x

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. https://doi.org/10.1016/j.physa.2009.05.033

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

Therneau, T., Atkinson, B., Ripley, B., & Ripley, M.B. (2018). Package 'rpart'. Retrieved on April 20, 2018 from cran.ma.ic.ac.uk/web/packages/rpart/rpart.pdf

Thrun, M.C. (2018). *Projection Based Clustering through Self-Organization and Swarm Intelligence* (A. Ultsch & E. Hüllermeier Adv.). Heidelberg, Germany: Springer. https://doi.org/10.1007/978-3-658-20540-9

Thrun, M.C., Lerch, F., Lötsch, J., & Ultsch, A. (2016). *Visualization and 3D Printing of Multivariate Data of Biomarkers*. In V. Skala (Ed.), International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision (WSCG), (pp. 7-16), Conference Proceedings. Plzen.

Thrun, M.C., & Ultsch, A. (2017). *Projection based Clustering.* Paper presented at the International Federation of Classification Societies (IFCS), Tokyo, Japan, August 7-10.

Thrun, M.C., & Ultsch, A. (2018). *Effects of the payout system of income taxes to municipalities in Germany*. In M. Papież & S. Śmiech (Eds.), 12th Professor Aleksander Zelias International Conference on Modelling and Forecasting of Socio-Economic Phenomena (pp. 533-542). Conference Proceedings. Cracow, Poland: Cracow: Foundation of the Cracow University of Economics.

Thrun, M.C., & Ultsch, A. (2019). *Analyzing the Fine Structure of Distributions. Technical Report*. Marburg, Germany: Philipps-University Marburg.

Ultsch, A. (2000). *Clustering with DataBots*.Int. Conf. Advances in Intelligent Systems Theory and Applications (AISTA) (pp. 99-104). Conference Proceedings. Canberra, Australia: IEEE ACT Section.

Ultsch, A. (2005). Pareto density estimation: A density estimation for knowledge discovery. In D. Baier & K.D. Werrnecke (Eds.), *Innovations in classification, data science, and information systems* (Vol. 27, pp. 91-100). Berlin: Springer.

Ultsch, A., & Lötsch, J. (2017). Machine-learned cluster identification in high-dimensional data. *Journal of Biomedical Informatics, 66*(C), 95-104. https://doi.org/10.1016/j.jbi.2016.12.011

Ultsch, A., & Thrun, M.C. (2017). *Credible Visualizations for Planar Projections*. In M. Cottrell (Ed.), 12th International Workshop on Self-Organizing Maps and Learning Vector Quantization, Clustering and Data Visualization (WSOM) (pp. 1-5). Conference Proceedings. Nany, France: IEEE.

Ultsch, A., Thrun, M.C., Hansen-Goos, O., & Lötsch, J. (2015). Identification of Molecular Fingerprints in Human Heat Pain Thresholds by Use of an Interactive Mixture Model R Toolbox (AdaptGauss). *International Journal of Molecular Sciences, 16*(10), 25897-25911. https://doi.org/10.3390/ijms161025897

UNDP. (2003). *Human development Report*. New York: In P. f. t. U. N. D. P. (UNDP) (Ed.).

Van der Maaten, L., & Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research, 9*(11), 2579-2605.

Vollmer, S., Holzmann, H., & Schwaiger, F. (2013). Peaks vs components. *Review of Development Economics, 17*(2), 352-364. https://doi.org/10.1111/rode.12036

Watanabe, S. (1969). *Knowing and Guessing: A Quantitative Study of Inference and Information*. New York, USA: John Wiley & Sons Inc.

Author

Michael C. Thrun

He graduated in with a diploma in physics in 2014 and received his PhD in Data Science from the Philipps University of Marburg in 2017. His research interests are centred around methods of dimensionality reduction, cluster analysis, data visualisation, as well as unsupervised machine learning with a specific focus on swarm intelligence, self-organisation and emergence. Currently, he works on such methods for advanced analytics in Big Data and generating industrial applications in data science.

Correspondence to: Michael C. Thrun, PhD, Philipps-University of Marburg, Hans-Meerwein-Straße 6, D-35032 Marburg, Germany, e-mail: mthrun@mathematik.uni-marburg.de

ORCID http://orcid.org/0000-0001-9542-5543

Acknowledgements and Financial Disclosure

Gratitude goes to Prof. Dr. Alfred Ultsch who made the World GDP dataset available. The author would like to express his gratitude to Prof. Dr. Andreas Geyer-Schulz for his valuable comments during the 12th Professor Aleksander Zelias International Conference on Modelling and Forecasting Socio-Economic Phenomena, where the draft of this paper was presented.

Copyright and License

Published by the Centre for Strategic and International Entrepreneurship – Krakow, Poland

Ministry of Science and Higher Education
Republic of Poland

The copyediting and proofreading of articles in English is financed in the framework of contract No. 845/P-DUN/2018 by the Ministry of Science and Higher Education of the Republic of Poland committed to activities aimed at science promotion.