Cluster Evolution Analytics

Morales-Oñate, Víctor and Morales-Oñate, Bolívar

Universidad de las Américas, Pontificia Universidad Católica del Ecuador sede Ambato

19 February 2024

Cluster Evolution Analytics

Víctor Morales-Oñate[a], Bolívar Morales-Oñate[b]

[a] *Universidad de las Américas, Departamento de Economía, Quito, Ecuador*
[b] *Pontificia Universidad Católica del Ecuador, Escuela de Ingenierías, Ambato, Ecuador*

Abstract

In this paper we propose Cluster Evolution Analytics (CEA) as a framework that can be considered in the realm of Advanced Exploratory Data Analysis or unsupervised learning. CEA leverages on the temporal component of panel data and it is based on combining two techniques that are usually not related: leave-one-out and plug-in principle. This allows us to use exploratory *what if* questions in the sense that the present information of an object is plugged-in a dataset in a previous time frame so that we can explore its evolution (and of its neighbors) to the present. We illustrate our results on a real dataset applying CEA on different clustering algorithms and developed a Shiny App with a particular configuration. Finally, we also provide an R package so that this framework can be used on different applications.

Keywords: clustering, temporal clustering, statistical profiles
2020 MSC: 91C20, 62H30

1. Introduction

Exploratory data analysis (EDA) shifted confirmatory data analysis to using data as the guiding principle to formulate hypothesis. The pioneering work of Tukey et al. (1977) is at the heart of this ongoing useful approach to Statistics. EDA leverages on context understanding, graphical representation (univariate, bivariate and multivariate), clustering, outlier detection, scaling, hypothesis suggestions, among others (Behrens, 1997).

Email address: `victor.morales@uv.cl` (Víctor Morales-Oñate)

Clustering in EDA has been used in the context of graphical methods, be it univariate, bivariate or multivariate data representation (Jebb et al., 2017). Multimodal distributions in univariate data signal the presence of more than one population. Scatterplots in bivaritate data suggest an underlying pattern of groups to be further understood. In multivaritate analysis there is usually some kind of dimension reduction before using graphical exploration to look for clusters in data. Chernoff faces and perceptual maps are examples multivaritate data exploration (Morris et al., 2000; Lee et al., 2016).

It is difficult to trace the roots of clustering and link it to a single author. Arguably, Sokal (1963, 1961) are some of the early works that tackle clustering with the name of taxonomy as is customary in life sciences. In these 60 years there have been rich advances in clustering, having established books (Xu & Wunsch, 2009; Kaufman & Rousseeuw, 2009; Everitt et al., 2011), fields such as Mutivariate Statistics (Harris, 2001; Izenman, 2008), Pattern Recognition (Ripley, 2007; Bishop & Nasrabadi, 2006) and ongoing research in unsupervised learning where clustering plays an important role (James et al., 2013; Hastie et al., 2009).

Detecting clusters in time has been tackled by several ways. One of them is by proposing a clustering index that accounts for a temporal clustering and the detection of cyclical clustering within a cycle length (Tango, 1984; Wallenstein, 1980). Another approach to clustering that takes time in consideration is in the context of data streaming. This problem is faced by finding clusters in data streams which may be frequent in time where scalability and functionality are some concerns (Aggarwal et al., 2003). Ezugwu et al. (2022) provide an up to date State-of-the-art survey of clustering in Machine Learning describing real world applications and techniques that are most widely used. Oliveira & Gama (2012) propose a framework to monitor the evolution of clusters: MEC. They emphasise the importance of taking into account the transitions of clusters over time and setup a taxonomy for that transition (birth, death, split, merge and survival) using a bipartite graph setup.

Despite the extensive literature studying clustering problems, to the best of our knowledge, there are no clear studies in exploratory data analysis or unsupervised learning that tackle clustering in time. Cluster evolution analytics lets the researcher propose exploratory *what if* questions in the sense of cluster evolution.

The reminder this paper is organized as follows. In Section 2 we revise the definition of clustering that CEA uses. In Section 3 we introduce the CEA framework and give a numerical example for its better understanding. In Section 4 we apply CEA framework to macroeconomic variables using Penn World Table 10.01 data source setting different scenarios of clustering algorithms. In Section 5 we describe the usage of a Shiny Application developed to a particular setting. In Section 6 we detail the main parameters used in CEA R package. In Section 7 we provide concluding remarks.

2. Clustering

A cluster is a grouping of objects that share similarities, and objects belonging to different clusters exhibit dissimilarities. Finding groups in data is the main objective of clustering. Clustering is partitioning an unlabeled finite dataset into a distinct set of underlying data structures that emerges from data (Kaufman & Rousseeuw, 2009).

We work with clustering in a hard partitioning setting. Following notation in Xu & Wunsch (2009), let $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_i, \ldots, \mathbf{x}_N\}$ be a set of input objects where $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ij}, \ldots, x_{id}) \in \mathbb{R}^d$, and x_{ij} is a feature (attribute, dimension or variable). A K-partition of $\mathbf{X}$, $C = \{C_1, \ldots, C_K\}$ $(K \leq N)$ is a hard partitioning if

1. $C_i \neq \emptyset$, $i = 1, \ldots, K$. Every cluster must have at least one element.
2. $\bigcup_{i=1}^{K} C_i = \mathbf{X}$. The union of all clusters is the input set $\mathbf{X}$.
3. $C_i \cap C_j = \emptyset$, $i, j = 1, \ldots, K$ and $i \neq j$. If an object belongs to a cluster, it cannot belong to another cluster.

Given the above definition of hard partitioning, it is also necessary to have proximity measures to assess how far (distance) or close (similarity) a pair of objects are (De Carvalho et al., 2012; Pfitzner et al., 2009). Another important metric is the linkage metric which let us measure the proximity between clusters (Murtagh & Contreras, 2012). Proximity measures between pais of objects and between clusters lets us compare different clustering algorithms.

A number of clustering algorithms have been derived in a hard partitioning context. K-means, K-medoids, DBSCAN, among others are some of the most widely used (Hubert & Arabie, 1985). CEA framework described in Section 3 can be used with any hard partition clustering algorith. Nonetheless, K-means, K-medoids are used in Sections 4 and Section 6.

3. Clustering Evolution Analytics

In what follows we introduce the cluster evolution analytics framework.

Let $\mathbf{X}^{t-l}$ be a set of input objects, $l \in \{0, 1, \ldots, T-1\}$ (T is the total time periods),

$$\mathbf{X}^{t-l} = \{\mathbf{x}_1^{t-l}, \mathbf{x}_2^{t-l}, \ldots, \mathbf{x}_i^{t-l}, \ldots, \mathbf{x}_N^{t-l}\}$$

where

$$\mathbf{x}_i^{t-l} = \left(x_{i1}^{t-l}, x_{i2}^{t-l}, \ldots, x_{ij}^{t-l}, \ldots, x_{id}^{t-l}\right) \in \mathbb{R}^d$$

with each x_{ij}^{t-l} is a feature.

Let $C^{t-l} = \{C_1^{t-l}, \ldots, C_{K_l}^{t-l}\}$ $(K \leq N)$ be a K partition of $\mathbf{X}^{t-l}$ at a fixed time l.

1. Select an object i from $\mathbf{x}_i^{t-l}$ to analyze its evolution.
2. Find its corresponding hard partition at $l = 0$, C^{t-0} and keep the neighbors of object i.
3. Remove $\mathbf{x}_i^{t-0}$ from $\mathbf{X}^{t-1}$.
4. Plug in the selected object $\mathbf{x}_i^{t-l}$ from step 1 in $\mathbf{X}^{t-1}$ such that,

$$\mathbf{X}_{-i}^{t-1} = \{\mathbf{x}_1^{t-1}, \mathbf{x}_2^{t-1}, \ldots, \mathbf{x}_i^{t-0}, \ldots, \mathbf{x}_N^{t-1}\}$$

.

5. Find C_{-i}^{t-1} being the hard partition of $\mathbf{X}_{-i}^{t-1}$ and keep the neighbors of object i
6. For $l \in \{2, \ldots, T-1\}$ and saving C_{-i}^{t-l}, repeat steps 3,4 and 5 until iteration T.

The output of the above steps is a list of neighbors of i where each element of the list has $K_0, K_1, \ldots, K_T$ neighbors (objects of the cluster that i belongs to) at every time l.

To illustrate CEA framework we consider a simple toy example and propose some questions that arise. Say we have a panel data as,

Time	Object	V1	V2
3	A	3	8
3	B	7	6
3	C	11	23
2	A	35	12
2	B	40	51
2	C	63	55
1	A	12	8
1	B	11	13
1	C	15	17

Following step 1, we select object $i = B$ and at time $t - 0 = 3$, we subset the dataset, obtain its partition and keep the neighbors of B at time 3, $NG_B^3 = \{A\}$ (step 2).

V1	V2
3	8
7	**6**
11	23

$$NG_B^3 = \{A\}$$

Now we remove $i = B$ from the subset $l = 2$ (step 3) and plug the values of $i = B$ from the subset $l = 3$ in the same location (step 4),

V1	V2
35	12
7	**6**
63	55

$NG_B^2 = \{C\}$

We now compute the hard partition and keep the neighbors of B at time 2, $NG_B^2 = \{C\}$ (step 5). Finally we keep iterating until $T-1$ (step 6). In our toy example $T=3$, so it stops at $l=2$,

V1	V2
12	8
7	**6**
15	17

$NG_B^1 = \{C\}$

The output is a list of neighbors NG_B (objects that belong to the same cluster as the selected i) of B for every l, $NG_B = \{A, C, C\}$. In our toy example they all have one neighbour but in general they can have different number of elements.

This simple example illustrates CEA framework. Note that some questions take place in light of these results: In general, today's B is similar to what objects in the past? What happened at time 2 so that C is no longer a neighbour of B at time 3? If C is at better conditions at time 3, what can we learn from C to replicate its success? If C is at worse conditions at time 3, what can we learn from C to avoid in the future? Of course, these questions are referential, other questions could be formulated depending on the researcher interests.

4. Application: Country economic profiles

It is impossible to pretend that an economic recipe is universally applicable in all countries. The heterogeneity that characterizes every nation is one of the main factors that makes such universality challenging. Economic convergence is a field of economics that studies questions such as: do automatic mechanisms exist that drive a convergence over time in per capita income and product levels between poor and rich nations? (Barro & Sala-i Martin, 1992). To answer this question, panel data and its associated econometrics methods are mostly used in empirical analysis (Barro, 1991; Bowdler & Malik, 2017; Durlauf et al., 2005; Sekrafi & Sghaier, 2016)

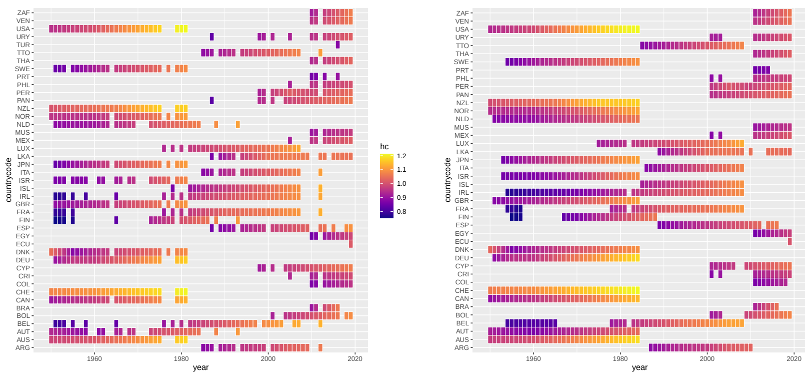
Using Unsupervised Machine Learning (ML) in Economics has caught recent attention between the research community. Athey & Imbens (2019) discuss Machine Leaning methods at the intersection of ML and econometrics and presents Text Analysis is one ways to exploit its intersection. CEA on its approach proposes a framework to analyse the cluster evolution of countries. Applying CEA in a yearly panel data of countries with macroeconomic variables can be summarized in the following steps:

1. Choose a country and a time range.
2. Detect clusters within a reference year and determine the cluster to which the chosen country belongs.
3. The data from the chosen country's base year is incorporated into the preceding time period.
4. Detect clusters in the previous time period and determine the cluster to which the chosen country belongs.
5. Iterate all time periods.

4.1. Data

This application uses data from Penn World Table (PWT). PWT version 10.01 is a database containing data on the relative levels of income, output, input, and productivity, spanning 183 countries from 1950 to 2019 (Feenstra et al., 2015). Table A.1 shows a description of the variables that can be found at PWT 10.01. Information is grouped in the following sections:

1. Real Gross domestic product (GDP), employment and population levels.
2. Current price GDP, capital and Total factor productivity (TFP).
3. National accounts-based variables.
4. Exchange rates and GDP price levels.



(a) (b)

Figure 1: CEA application for Ecuador between 1950-2019 with k-means algorithm (1a) k-medoids (1b)

5. Shares in Current Price Gross domestic Product Output-side (CGDPo).
6. Price levels, expenditure categories and capital.

A complete panel data is obtained for 53 countries from 1950 to 2019 for Human capital index (`hp`). In our application, we select Ecuador for the analysis.

4.2. CEA framework applied

In this particular application of CEA, we use Ecuador as the selected country and study Human capital index (`hp`) . Figure 1 shows a heatmap of CEA results of 53 countries in 1950-2019. The left panel applies CEA using k-means algorithm and the right panel uses k-medoids (partitioning around medoids) algorithm. In a general sense, both algorithms tend to cluster the same countries. Nonetheless, k-means is more sensitive to turn *on* and *off* Ecuador's neighbors in the period. For example, while Denmark is consistently Ecuador's neighbour in 1950-1984, in the same period k-means do not cluster them together in 1955, 1957, 1962, 1977, 1978, 1982, 1983 and 1984.

Figure 2 shows the evolution of the number of Ecuador's neighbors in the period. Solid line shows k-means method and the dotted line shows the k-medoids method. This result also confirms that k-means is more volatile than k-medoids.

Figure 2: Evolution of the number of neighbors of Ecuador.

Recalling questions proposed in Section 3, lets use this macroeconomic application to answer them:

- In general, today's Ecuador is similar to what countries in the past? Using k-medoids algorithm from 1950 to 2019, a possible answer is listing the most frequent countries that are grouped along with Ecuador. They are: United States, Belgium and Ireland.

- What happened at in the time range so that Belgium is no longer a neighbour of Ecuador in 2019? Note that Ecuador being similar to developed countries in Human Capital may be surprising. However, Figure 3 shows values of Human Capital of Belgium and Colombia as Ecuador's neighbors in time. Ecuador's Human Capital in 2019 is 1.016 (log scaled) has similar values from 1978 (0.924) to 2008 (1.126). Belgium has an increasing trend s that after 2008 it no longer is in Ecuador's group.

- If Belgium is at better conditions at 2019, what can we learn from Belgium to replicate its success? In 2019, Belgium's Human Capital is 3.15, its development is far from Ecuador's.

- If Colombia is at worse conditions in 2019, what can we learn from Colombia to avoid in the future? Colombia's Human Capital in 2019 is 0.956.

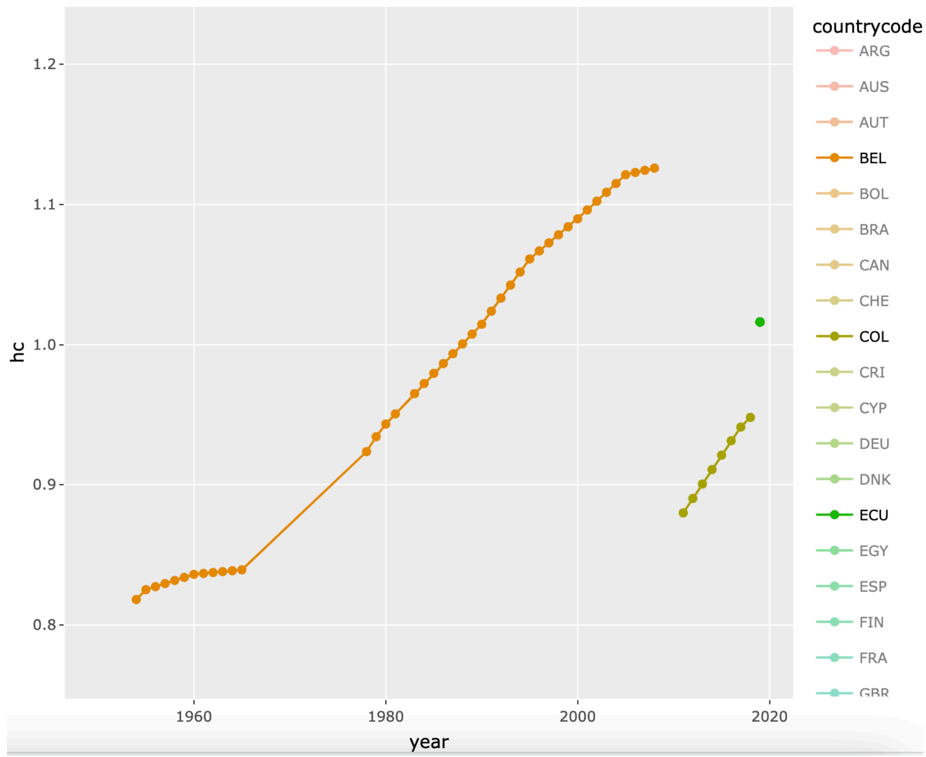


Figure 3: Evolution Belgium and Colombia as Ecuador's neighbors.

> Figure 3 shows that Colombia is in Ecuador's group from 2011 which means that it is similar to Ecuador's 2019 Human Capital in recent years. Ecuador should look close to Colombia's increasing trend since it will probably leave behind Ecuador as Belgium did.

Note that the proposed questions in Section 3 were hypothetical and in the application they can be answered using data as exploratory *what if* questions.

5. Shiny App: Country Macroeconomic profiles

In order to get more insight of the potential of CEA applications, we developed a Shiny App that applies CEA to Country macroeconomic profiles that can be found at `https://vmoprojs.shinyapps.io/ClusEvol`. It is a complement of Section4 and also lets the researcher choose different parameters:

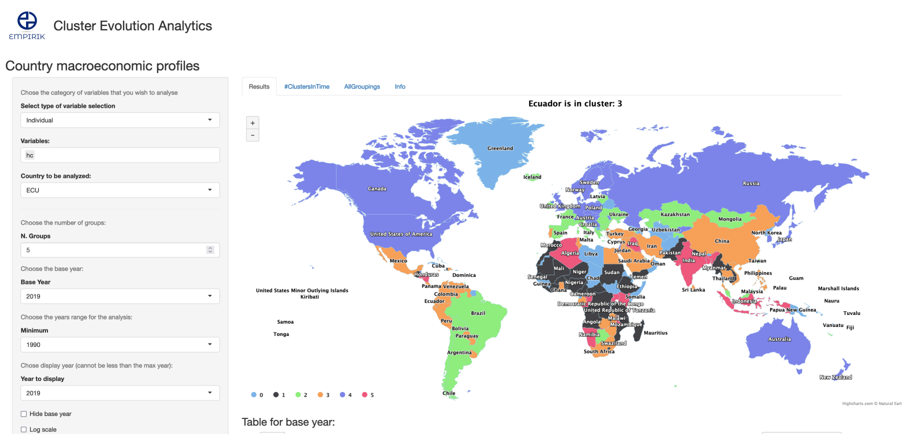


Figure 4: Cluster Evolution Analytics-Country macroeconomic profiles Shiny App front.

- *Select type of variable selection* sets the option that the user chooses to use grouped variables or select variables one by one.
- *Variables* lets the user select individual variables if *Select type of variable selection* is not grouped.
- *Country to be analyzed* lets the user choose the country to be analyzed.
- *N. Groups* sets the number of groups to be used in k-means clustering.
- *Base Year* sets the initial year of the time range. It usually is the maximum of the time range.
- *Minimum* sets the initial year of the time range. It usually is the minimum of the time range.
- *Year to display* sets the year that is shown in Figure 4.
- *Hide base year*. It hides the table under the map if checked.
- *Log scale* it log transform input variables for k-means algotirhm if checked.

The application lets the user change the listed parameters and results are presented in different panels. The source code of the Shiny Application can be found at `https://github.com/vmoprojs/ShinyApps/tree/master/ClusEvol`.

6. CEA: the package `clusEvol`

Besides the application described in Section 4 and its Shiny version in Section 5, Cluster Evolution Analytics (CEA) can be used in several fields that the researcher finds it useful. This motivates the development of an R package that lets the user apply CEA: clusEvol package. The user can install the package with the following code:

```
devtools::install_github("https://github.com/vmoprojs/clusEvol").
```

`clusEvol` contains a panel dataset (`actpas`) of Ecuador's amount of Assets and Liabilities Operations of the National Financial System. The main function of the package has the same name as the package. The following code results a CEA application to `actpas`:

```
library(clusEvol)
data(actpas)
solclusEvol <- clusEvol(x=actpas,objects="razon_social",
time = "fecha",target.vars = c("montoAct","operAct"),
                              time.base=max(actpas$fecha),
                              sel.obj="BANCO SOLIDARIO S.A.",
                              init = min(actpas$fecha),
                              logscale = TRUE,ng = 5,clm = "pam")
print(solclusEvol)
```

A detailed description of `clusEvol` parameters can be found by `help(clusEvol)`. The `print` method gives information about

- Number of neighbors `sel.obj` is a group member
- Cluster that `sel.obj` belongs to
- Clusters in time.

The package also have a `plot` method by which Figure 1 was obtained. Finally, other panel datasets can be used. For example, Grunfeld panel data from `plm` (Croissant & Millo, 2008) is used in the following code:

```r
data("Grunfeld", package="plm")
library(clusEvol)
sel.obj <- "8"
solclusEvol <- clusEvol(x=Grunfeld,objects="firm",
                time = "year",
                target.vars = c("inv","value","capital"),
                time.base=1954,sel.obj="8",init = 1935,
                logscale = TRUE,
                ng = 5,clm = "pam",scale = FALSE)
```

`clusEvol` can be applied to datasets with a panel data structure. Interpretations of the results will vary depending on the specific application and the researcher's expertise in the field.

7. Results and discussion

The Cluster Evolution Analytics (CEA) framework is introduced as a tool for proposing and gaining insights into *what if* scenarios using data. This article discusses various applications of CEA that can assist researchers in exploring field-specific questions, provided they have access to panel data.

Offering both a Shiny Application and an R package to extend the utilization of the CEA framework can enhance researchers' understanding of their particular applications. This facilitates a fresh perspective on data analysis, incorporating not just observational units but also temporal considerations. Nonetheless, it's important to acknowledge the limitations of CEA. For instance, it's imperative to recognize that CEA does not inherently address causality.

The current research lays a foundation for further exploration and development. Diving into its various facets could yield enhanced versions of CEA. Possibilities include integrating CEA with methods for *optimal* cluster number selection, investigating additional hard partition algorithms, or even formulating a CEA variant tailored for fuzzy clustering.

Just like in Exploratory Data Analysis, the researcher's domain expertise remains crucial in the application of CEA. However, even individuals new to statistics can benefit from the framework. With the support of an R package, conducting data exploration through CEA can unlock valuable insights for newcomers, empowering them to extract meaningful information from their data.

It's important to note that CEA occupies a place within Unsupervised Learning, offering accessibility across a spectrum of scientific disciplines, ranging from the natural sciences to the social sciences. This versatility underscores its potential to contribute valuable insights across diverse fields of study.

References

Aggarwal, C. C., Philip, S. Y., Han, J., & Wang, J. (2003). A framework for clustering evolving data streams. In *Proceedings 2003 VLDB conference* (pp. 81–92). Elsevier. doi:`https://doi.org/10.1016/B978-012722442-8/50016-1`.

Athey, S., & Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics*, *11*, 685–725. doi:`https://doi.org/10.1146/annurev-economics-080217-053433`.

Barro, R. J. (1991). Economic growth in a cross section of countries. *The quarterly journal of economics*, *106*, 407–443.

Barro, R. J., & Sala-i Martin, X. (1992). Convergence. *Journal of political Economy*, *100*, 223–251.

Behrens, J. T. (1997). Principles and procedures of exploratory data analysis. *Psychological methods*, *2*, 131.

Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning* volume 4. Springer.

Bowdler, C., & Malik, A. (2017). Openness and inflation volatility: Panel data evidence. *The North American Journal of Economics and Finance*, *41*, 57–69.

Croissant, Y., & Millo, G. (2008). Panel data econometrics in R: The plm package. *Journal of Statistical Software*, *27*, 1–43. doi:`10.18637/jss.v027.i02`.

De Carvalho, F. D. A., Lechevallier, Y., & De Melo, F. M. (2012). Partitioning hard clustering algorithms based on multiple dissimilarity matrices. *Pattern Recognition*, *45*, 447–464. doi:`https://doi.org/10.1016/j.patcog.2011.05.016`.

Durlauf, S. N., Johnson, P. A., & Temple, J. R. (2005). Growth econometrics. *Handbook of economic growth*, *1*, 555–677.

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster analysis*. John Wiley & Sons.

Ezugwu, A. E., Ikotun, A. M., Oyelade, O. O., Abualigah, L., Agushaka, J. O., Eke, C. I., & Akinyelu, A. A. (2022). A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. *Engineering Applications of Artificial Intelligence*, *110*, 104743. doi:https://doi.org/10.1016/j.engappai.2022.104743.

Feenstra, R. C., Inklaar, R., & Timmer, M. P. (2015). The next generation of the penn world table. *American economic review*, *105*, 3150–3182. doi:https://doi.org/10.34894/QT5BCC.

Harris, R. J. (2001). *A primer of multivariate statistics*. Psychology Press.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* volume 2. Springer.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of classification*, *2*, 193–218. doi:https://doi.org/10.1007/BF01908075.

Izenman, A. J. (2008). *Modern multivariate statistical techniques* volume 1. Springer.

James, G., Witten, D., Hastie, T., Tibshirani, R. et al. (2013). *An introduction to statistical learning* volume 112. Springer.

Jebb, A. T., Parrigon, S., & Woo, S. E. (2017). Exploratory data analysis as a foundation of inductive research. *Human Resource Management Review*, *27*, 265–276.

Kaufman, L., & Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons.

Lee, A. J., Yang, F.-C., Chen, C.-H., Wang, C.-S., & Sun, C.-Y. (2016). Mining perceptual maps from consumer reviews. *Decision Support Systems*, *82*, 12–25. doi:https://doi.org/10.1016/j.dss.2015.11.002.

Morris, C. J., Ebert, D. S., & Rheingans, P. L. (2000). Experimental analysis of the effectiveness of features in chernoff faces. In *28th AIPR Workshop: 3D Visualization for Data Exploration and Decision Making* (pp. 12–17). SPIE volume 3905. doi:https://doi.org/10.1117/12.384865.

Murtagh, F., & Contreras, P. (2012). Algorithms for hierarchical clustering: an overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *2*, 86–97. doi:https://doi.org/10.1002/widm.53.

Oliveira, M., & Gama, J. (2012). A framework to monitor clusters evolution applied to economy and finance problems. *Intelligent Data Analysis*, *16*, 93–111. doi:10.3233/IDA-2011-0512.

Pfitzner, D., Leibbrandt, R., & Powers, D. (2009). Characterization and evaluation of similarity measures for pairs of clusterings. *Knowledge and Information Systems*, *19*, 361–394. doi:https://doi.org/10.1007/s10115-008-0150-6.

Ripley, B. D. (2007). *Pattern recognition and neural networks*. Cambridge university press.

Sekrafi, H., & Sghaier, A. (2016). Examining the relationship between corruption, economic growth, environmental degradation, and energy consumption: a panel analysis in mena region. *Journal of the Knowledge Economy*, (pp. 1–17).

Sokal, R. R. (1961). Distance as a measure of taxonomic similarity. *Systematic Zoology*, *10*, 70–79.

Sokal, R. R. (1963). The principles and practice of numerical taxonomy. *Taxon*, (pp. 190–199).

Tango, T. (1984). The detection of disease clustering in time. *Biometrics*, (pp. 15–26). doi:https://doi.org/10.2307/2530740.

Tukey, J. W. et al. (1977). *Exploratory data analysis* volume 2. Reading, MA.

Wallenstein, S. (1980). A test for detection of clustering over time. *American Journal of Epidemiology*, *111*, 367–372. doi:https://doi.org/10.1093/oxfordjournals.aje.a112908.

Xu, R., & Wunsch, D. (2009). *Clustering*. John Wiley & Sons.

Appendix A. Data description

Real Gross domestic product (GDP), employment and population levels	
rgdpe	Expenditure-side real GDP at chained Purchasing power parities (PPPs) (in mil. 2017US$)
rgdpo	Output-side real GDP at chained PPPs (in mil. 2017US$)
pop	Population (in millions)
emp	Number of persons engaged (in millions)
avh	Average annual hours worked by persons engaged
hc	Human capital index, based on years of schooling and returns to education; see Human capital in PWT9.
Current price GDP, capital and Total factor productivity (TFP)	
ccon	Real consumption of households and government, at current PPPs (in mil. 2017US$)
cda	Real domestic absorption, (real consumption plus investment), at current PPPs (in mil. 2017US$)
cgdpe	Expenditure-side real GDP at current PPPs (in mil. 2017US$)
cgdpo	Output-side real GDP at current PPPs (in mil. 2017US$)
cn	Capital stock at current PPPs (in mil. 2017US$)
ck	Capital services levels at current PPPs (USA=1)
ctfp	TFP level at current PPPs (USA=1)
cwtfp	Welfare-relevant TFP levels at current PPPs (USA=1)
National accounts-based variables	
rgdpna	Real GDP at constant 2017 national prices (in mil. 2017US$)
rconna	Real consumption at constant 2017 national prices (in mil. 2017US$)
rdana	Real domestic absorption at constant 2017 national prices (in mil. 2017US$)
rnna	Capital stock at constant 2017 national prices (in mil. 2017US$)
rkna	Capital services at constant 2017 national prices (2017=1)
rtfpna	TFP at constant national prices (2017=1)
rwtfpna	Welfare-relevant TFP at constant national prices (2017=1)
labsh	Share of labour compensation in GDP at current national prices
irr	Real internal rate of return
delta	Average depreciation rate of the capital stock
Exchange rates and GDP price levels	
xr	Exchange rate, national currency/USD (market+estimated)
pl_con	Price level of CCON (PPP/XR), price level of USA GDPo in 2017=1
pl_da	Price level of CDA (PPP/XR), price level of USA GDPo in 2017=1
pl_gdpo	Price level of CGDPo (PPP/XR), price level of USA GDPo in 2017=1
Shares in CGDPo	
csh_c	Share of household consumption at current PPPs
csh_i	Share of gross capital formation at current PPPs
csh_g	Share of government consumption at current PPPs
csh_x	Share of merchandise exports at current PPPs
csh_m	Share of merchandise imports at current PPPs
csh_r	Share of residual trade and GDP statistical discrepancy at current PPPs
Price levels, expenditure categories and capital	
pl_c	Price level of household consumption, price level of USA GDPo in 2017=1
pl_i	Price level of capital formation, price level of USA GDPo in 2017=1
pl_g	Price level of government consumption, price level of USA GDPo in 2017=1
pl_x	Price level of exports, price level of USA GDPo in 2017=1
pl_m	Price level of imports, price level of USA GDPo in 2017=1
pl_n	Price level of the capital stock, price level of USA in 2017=1
pl_k	Price level of the capital services, price level of USA=1

Table A.1: Variables in Penn World Table (PWT) Version 10.01