A New Method of Identifying Key Industries: A Principal Component Analysis

Lefteris Tsoulfidis (Lnt@uom.edu.gr)
University of Macedonia https://orcid.org/0000-0003-2691-0128
Ioannis Athanasiadis
University of Macedonia

Research

Keywords: Principal components, structural change, dimensionality reduction, clusters, networks

Posted Date: November 22nd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1053053/v1

A New Method of Identifying Key Industries: A Principal Component Analysis

By

Lefteris Tsoulfidis* and Ioannis Athanasiadis**

ABSTRACT

This article using the principal components analysis identifies key industries and groups them into particular clusters. The data come from the US benchmark input-output tables of the years 2002, 2007, 2012 and the most recently published input-output table of the year 2019. We observe some intertemporal switches of industries both between and within the top clusters. The findings further suggest that structural change is a slow moving process and it takes time for some industries to move from one cluster to the other. This information may be proved important in the designation of effective economic policies by targeting particular industries and also for the stability properties of the economic system.

JEL classifications: B24, B51, C67, D46, D57, E11, E32

Key Words: Principal components, structural change, dimensionality reduction, clusters, networks

* Department of Economics, University of Macedonia, Thessaloniki, Greece, Lnt@uom.edu.gr

** Department of Economics, University of Macedonia, Thessaloniki, Greece, athang@uom.edu.gr

1. Introduction

In this article, we apply dimensionality reduction to three benchmark input-output tables of the USA of the years 2002, 2007 and 2012 as well as the last available input-output table of the year 2019.[1] The dimensions of the tables are reduced to 70x70, as we have eliminated the industry housing because of its many imputations, and also the fictitious household industry. The idea of dimensionality reduction, that is, the way to reduce the complexity in the modeling of an economy can be traced in the writings of the Physiocrats and their tableau économique, whose purpose was to mimic the operation of the entire economy by compressing it into three sectors. We do know that the tableau is essentially the prototype of a one (multipurpose) commodity world. A similar idea can be discerned in Ricardo's corn model and certainly in Marx's schemes of simple reproduction, where there is a single commodity that functions either as a consumer or an investment good. As this can be judged by the uniform capital-intensity in the two sectors (departments) of the economy. A similar idea was advanced by Samuelson (1962) with his parable production function also based on the production of a single commodity. Finally, in recent economic growth theory, we increasingly observe the case of models of a one-commodity world.

In the works (Mariolis and Tsoulfidis 2018, Tsoulfidis 2021) there has been an effort to stripe down the behavior of the entire economic system and compress it into a single hyper-industry through the application of the Schur and singular value decomposition (SVD) techniques (Meyer 2002, ch. 5). The rationale for the application of these techniques is the particular skew distribution of the economic system's eigenvalues and the distant to the maximal, second along with the subdominant eigenvalues, whose impact on the economy is minimal and for all practical purposes can be side-stepped. Similarly, using the principal components analysis (PCA), we separate the impact of the top two (at most three) eigenvalues, which is equivalent to saying that the movement of prices induced by changes in the rate of profit is curvilinear and the same is true with the wage rate of profit curves or what is the same factor price frontier. Such findings suggest that the distribution of eigenvalues is mainly responsible for the shape of the price-profit or wage-profit curves and that the first eigenvalue along with the second compress most of the information regulating the economic system's motion leaving not much to be explained by the third or fourth eigenvalues (Tsoulfidis 2021, ch. 6).

The remainder of the article continues as follows: Section 2 gives a brief description of the fundamentals of the PCA and its use in determining clusters in the economy. Section 3 applies the PCA and contrasts the first principal component against the backward linkages of the industries. Section 4 introduces the k-means clustering technique to extract the optimal number of groups of industries and then using the silhouette method forms the clusters of industries. Section 5 orders each of the

[1] The data were downloaded in March 2021.

industries in the form of a dendrogram and respective network. Finally, Section 6 summarizes and concludes with remarks about future directions of the research.

2. Methods and Results

2.1 Dimensionality Reduction through Principal Components Analysis

PCA is an effective dimensionality reduction technique that constructs relevant features through linear combinations of the original features. The construction of relevant features is carried out by linearly transforming correlated variables into fewer uncorrelated variables. This transformation becomes possible through the projection of the initial data into the reduced PCA space using the eigenvectors of the covariance/correlation matrix, or what is the same, the principal components (PCs). The resulting projected data comprise essentially linear combinations of the initial data capturing most, if not all, of the variance in the data. Furthermore, the PCA increases interpretability and, in so doing, can become particularly helpful in dealing with economic datasets. Notwithstanding its great advantages PCA, to the extent, we know the literature, has not been used in major economic questions except in the areas of finance (e.g., Plerou *et al*. 2002, Farné and Vouldis 2021).

The advantages of the PCA renders it particularly applicable in input-output data in the direction of identifying the relative importance of industries in the operation of the economy. From the estimated PC the dimensionality reduction is restricted to the top two, which are supposed to compress most of the influence or what is the same most of the variance in the data. A third or a fourth PC could also be included, if such a need arises, but then the model loses its parsimonious character, and its analytical strength gradually fades away. In dealing with the available input-output data and their eigendecomposition, we have repeatedly found that the linear and quadratic approximation of the movement of prices induced by changes in income distribution is pretty accurate. The remaining terms from the eigendecomposition, although many add only minuscule information and we can dispense with it (see Bienenfeld 1989, Iliadi et al. 2014, Mariolis and Tsoulfidis 2018, Tsoulfidis 2021, ch. 6). The application of the PC analysis in our input-output data composed of the following main operation steps:

Step 1
Starting with the matrix of direct input requirements or Leontief inverse provided by the BEA for the benchmark years, that is, the matrix of total requirements or Leontief inverse $[\mathbf{I}-\mathbf{A}]^{-1}$, we estimate the input-output matrix

$$\mathbf{A}=\mathbf{I}-([\mathbf{I}-\mathbf{A}]^{-1})^{-1}$$

From which in turn, we get the vertically integrated input-output coefficients,

$$\mathbf{H} = \mathbf{A}[\mathbf{I} - \mathbf{A}]^{-1}$$

and so, we end up with the matrix $\mathbf{H}$ of 70x70 industries, whose PC we seek to estimate in the effort to group industries into clusters of different importance and meaning. The advantage of the matrix $\mathbf{H}$ is that it gives more (less) weight to the larger (smaller) coefficients in matrix $\mathbf{A}$.[2] Furthermore the matrix $\mathbf{H}$ is used in the estimation of prices and their changes in the face of income redistribution (Tsoulfidis 2021, ch. 6 and the literature cited there).

Step 2

We centered the matrix $\mathbf{H}$, by subtracting from each column element the column mean and we repeat the process for each of our 70 industries and get

$$\bar{\mathbf{H}} = \mathbf{H} - \mathbf{e}'\mathbf{e}\mathbf{H}/70$$

Where $\mathbf{e}$ is the row (1x70) unit or summation vector. In order to find the variance/covariance matrix, we multiply $\bar{\mathbf{H}}$ from the left by $\bar{\mathbf{H}}'$ and form the new matrix

$$\bar{\mathbf{H}}'\bar{\mathbf{H}}/(70-1)$$

the eigenvalues of the above variance/covariance matrix ranked from the maximum to minimum along with their respective eigenvectors. It is important to note that the eigenvalues of the matrix $\bar{\mathbf{H}}'\bar{\mathbf{H}}/69$, the maximal eigenvalue stands for the maximal variance and so forth for the rest.

Step 3

We know that the eigenvalues denote the relative importance of their corresponding eigenvectors. It follows, therefore, that the ratio of each of the eigenvalues relative to their total sum gives the proportion of variance explained. From the estimated eigenvalues, we isolate the first couple, whose percentage in the total is found quite satisfactory. This requirement has indicated that with the first two eigenvalues, despite their relatively low percentage, they provide a pretty accurate description of the motion of the entire economic system. As a consequence, by adding the next in ranking eigenvalues, we do not improve our overall understanding of the relative importance and interconnections of industries. The eigenvectors indicate both the size and the direction of the variance and they are ranked according to their respective eigenvalues starting with the maximal going to the second, third, and so forth. We rotate the

[2] The Leontief inverse of the input-output matrix accounts for both the direct and indirect interindustry linkages.

eigenvectors such that to place the first PC on the horizontal axis and the second PC on the vertical axis.

Figuratively speaking, and in case of two PC, we may think of the first PC as the orthogonal that one gets from the vertical intersection of a cylinder, provided that its height is by far higher than its periphery. The second PC is a horizontal intersection of the same cylinder perpendicular to the orthogonal. The variance equated with the height of the first PC is meaningful only if it is significantly larger than the second PC and both are distant enough from the rest. In our data a third PC will give rise to a three dimensional graph but this would not add much in our denoising process and the extraction the relative importance of each of our 70 industries. The PCA seeks to maximize, to the extent possible, the information content in the first PC, the remaining information is in the second PC and so on. The scree plot, that is, the distribution of eigenvalues of the matrix $\overline{\mathbf{H}}'\overline{\mathbf{H}}/69$, signifies the relative importance of each of the PC.

Step 4

Having selected the first two eigenvalues of the variance/covariance matrix $\overline{\mathbf{H}}'\overline{\mathbf{H}}/69$ and respective eigenvectors or rather feature vectors discarding those of lesser significance (of lower than the second eigenvalues), and form with the remaining ones a matrix of vectors that we call feature vector. In short, the feature vectors simply form a matrix that has as columns the eigenvectors of the components. This makes it the first step towards dimensionality reduction, because if we choose to keep only the first two eigenvectors (components) out of 70, the final data set will have only 70x2 dimensions. Subsequently, for the derivation of PC we apply the following multiplication

$$\overline{\mathbf{PC}} = (\overline{\mathbf{H}}'\overline{\mathbf{H}}/69)\mathbf{PC}$$

and we end up with the $\overline{\mathbf{PC}}$ of the economic system matrices from which we keep the eigenvectors corresponding to the top two eigenvalues.

2.2 Principal Components and Clusters of Industries with Input-Output Data

Before we introduce the PCA and its application in identifying key industries in the input-output structure of the economy it is important to establish its connection and relation in general with what has been hitherto used in input-output analysis. For this purpose, we start with the Leontief inverse, which is directly given in the input-output tables of the BEA, and make a comparison between the estimated first PC and the estimates of linkages, backward or forward. From the examination of the simple additions of columns or rows of total requirement matrices of the benchmark years 2002, 2007, 2012 and also 2019, the last input-output data available, we find as expected that the simple row sum of $[\mathbf{I} - \mathbf{A}]^{-1}$ has a low correlation with the

estimated first PC of the matrix $\mathbf{H}$. By contrast the column sums or the column norms of the matrix $[\mathbf{I}-\mathbf{A}]^{-1}$ are closely associated with the first PC.

It is important to emphasize at this juncture that the forward linkages (FL) or backward linkages (BL) of industries of matrices have the exact same ranking regardless of the use of matrix $[\mathbf{I}-\mathbf{A}]^{-1}$ or the vertically integrating technical coefficients matrix $\mathbf{H}$, which is used for the estimation of our PC. We opted for the estimation of PC to utilize the matrix $\mathbf{H}$ on the basis of which the estimation of relative prices and wage rate of profit curves or price factor frontiers are carried out (Tsoulfidis 2021). The estimated first PC from the above matrix is found to be highly correlated with the direct (unweighted) BL of the matrix $(\mathbf{e}[\mathbf{I}-\mathbf{A}]^{-1})'/70$, where $\mathbf{e}$ is the row (1x70) summation vector of ones, the so derived industry average backward linkage was further divided by the economy's average (Watanabe and Chenery 1958, Miller and Blair 2009, ch.12). If an industry's linkages are higher than one, it follows that this particular industry weighs more than the economy-wide average. As a consequence, changes in this particular industry exert higher than average effects on the total economy, the converse holds for those with less than one.

The first PC corresponding to the eigenvector with the maximal eigenvalue provides us with information as to where the data is maximally spread out and, therefore, explains the most variance of the system. The second PC has a lower eigenvalue and thus encompasses most of the system's remaining variance. The skew distribution of eigenvalues allows the selection of the top two eigenvalues, which although they account for nearly 50% of the total variance; nevertheless, are adequate since we are dealing with too many (70x70) observations.

The idea is that an industry with a high power of dispersion (variance) concentrates the features of a key industry. This is equivalent to saying that a given percentage increase in its output will deliver a significant impact on its suppliers. By contrast, in an industry with relatively small variance, a change in its output draws evenly and lightly on its suppliers.[3]

It is also important to note that we did not get an equally strong relationship between the principal components of the matrix $\mathbf{H}$ and the FL. The idea is that for the sum of rows, we refer to output proportions and so a closer relationship would require estimation of $\overline{\mathbf{PC}}$ from the matrix $\mathbf{H}' = [\mathbf{I}-\mathbf{A}]^{-1}\mathbf{A}$. This does not mean that the FL are not important in the understanding of interrelationships between industries and the structural changes in the economic system as a whole. For this purpose neither the first PC is adequate and needs to be supplemented by the second, at least, PC.

[3] In the literature there various ways to weigh the relative importance of each industry, for example, relative industries' shares in final demand or total output instead of unweighted measures as the one we used. However, we do not expect any qualitative differences in our results.

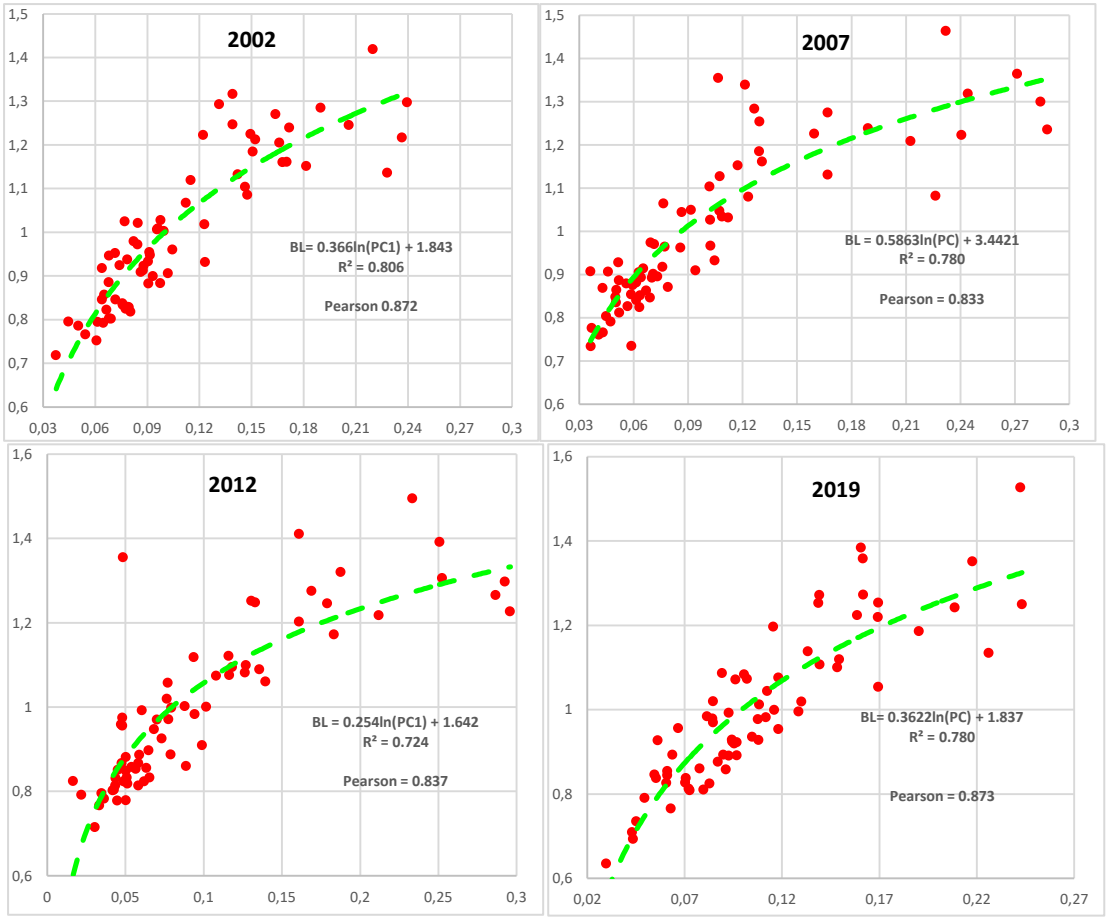


Figure 1. Principal components vs. backward linkages a horn kind of relationship

From the panel of four graphs (Figure 1) we observe that the first PC is strongly associated with the BL of each industry, as this is reflected in the high R-square, which varies from 72 to 80 percent. The Pearson correlation coefficient (displayed in each of the panel of four graphs in Figure 1), which is well above 0.80 indicates a strong similarity and positive relationship between the two in comparison variables, in each year of our study. These findings suggest that the application of the PCA may lead to fuller and more informative results with respect to the interrelations of industries. Furthermore, since we are looking for higher variances, it becomes particularly operational in evaluating the relative importance of each industry, as one may judge by the covariance of the central array and then find the eigenvalue and eigenvectors of the covariance. The dominant eigenvalue explains most of the variance in the data and the second along the subdominant eigenvalues ranked from the top down are used to categorize each industry to appropriate cluster as shown below. The covariance matrix $\overline{\mathbf{H}}'\overline{\mathbf{H}}/69$ gave the following eigenvalues for each of the four years of our study, which we plot in a panel of scree graphs in Figure 2 below

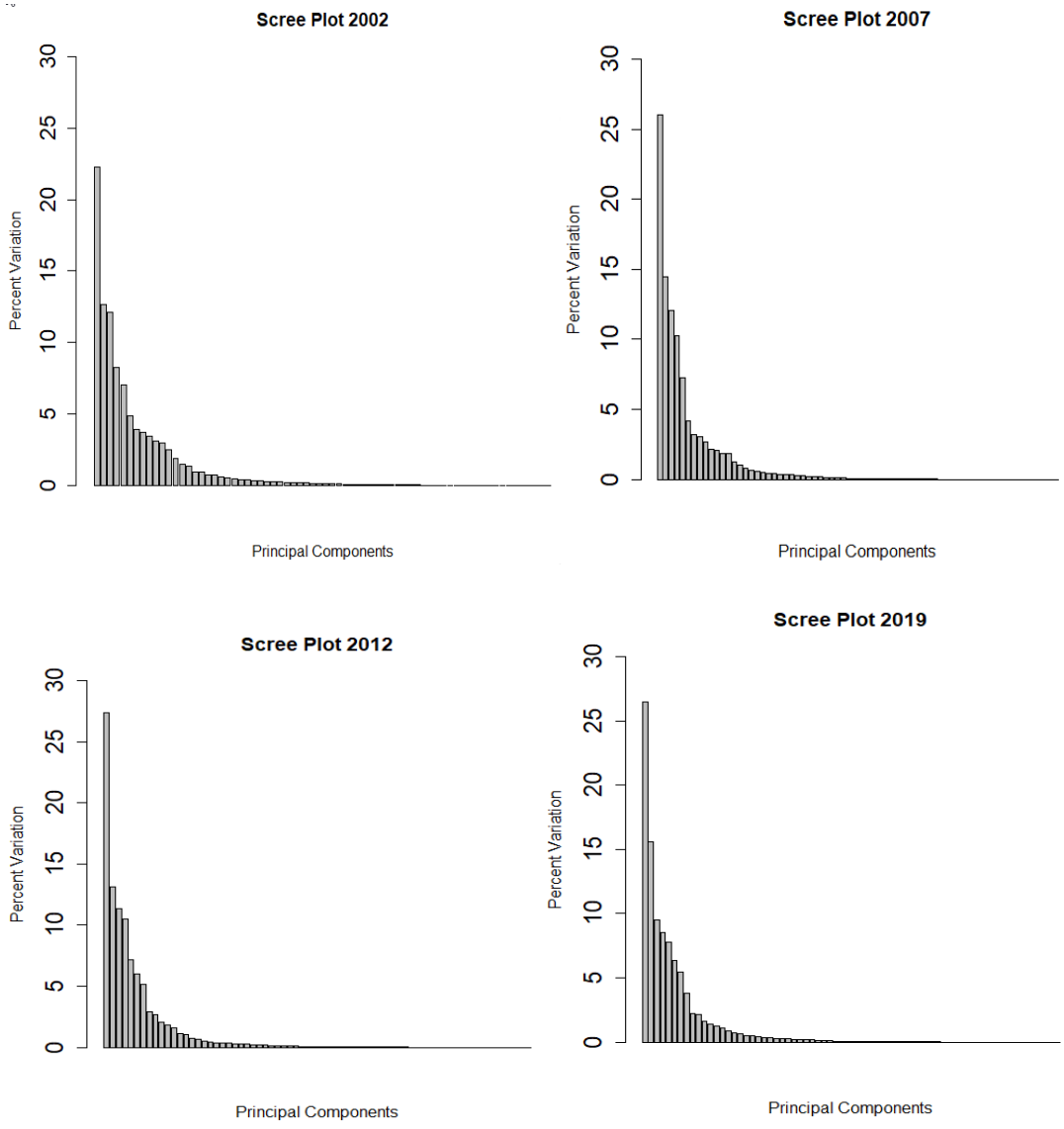


Figure 2. Scree plots of covariance matrices 2002, 2007, 2012, and 2019

In spectral methods, the top eigenvalues decide on the dimensions or, what is the same, the number of PC to be selected depending on the so-called "elbow rule."

And the elbow rule by and large dictates that the percentage explained by the PC should exceed 70%, but when we are dealing with large dimensions, as in our case, this percentage might be significantly lower. From the panel of four scree graphs in Figure 2, we observe that the first two (or at most three) eigenvalues are much higher than the rest and from the third eigenvalue onwards starts the decaying of eigenvalues. By adding the subdominant eigenvalues and the associated with these eigenvectors, we do not improve our overall explanation or variability, and certainly, we do not affect qualitatively our results.

2.3 Identifying Industry Clusters

In data science and especially in datasets with many features (or variables) such as in our case, clustering is a very useful tool. The greatest dissimilarity between different clusters and most important the greatest similarity within the same cluster, is the method for finding cluster structure in a dataset. The earliest method used mainly by biologists and social scientists is the hierarchical clustering, setting cluster analysis as a branch of multivariate statistical analysis (Jain and Dubes 1988, Kaufman and Rousseeuw 1990). This approach to machine learning is also called unsupervised learning.

These clustering methods, from a statistical view point, generally are considered as probability and non-parametric model-based approaches. The first approach follow that the observations are from a mixture probability model with the consequence to use a mixture likelihood approach to clustering (McLachlan and Basford, 1988). The Expectation and Maximization (EM) algorithm is the most frequently used in model-based approaches (Dempster et al. 1977, Yu et al. 2018). The use of an objective function of similarity or dissimilarity measure is the preferred method for clustering in a non-parametric approach, with the division into hierarchical and partitional methods in nearly every case (Kaufman and Rousseeuw 1990, Jain 2010, Yang et al. 2018).

On the other hand, the main problem of these k-means clustering algorithms is the need to give a number of clusters a priori. For solving this, validity indices supposed to be independent of clustering algorithms should be used (Halkidi et al. 2001). Many of these validity indices had been proposed such as Bayesian Information Criterion (BIC) (Kass and Raftery, 1995), Akaike Information Criterion(AIC) (Bozdogan, 1987), Dunn's index (Dunn, 1973), Davies-Bouldin index (DB) (Davies and Bouldin, 1979), Silhouette Width (SW) (Rousseeuw, 1987) and Gap statistic (Tibshirani et al., 2001) among others.

How similar an object is to its own cluster (cohesion) compared to other clusters (separation) is a very useful measure obtained by the Silhouette value (Rousseeuw, 1987) that we use in our approach in deriving the key industries of the economy. This measure ranges from -1 to +1 with higher values better match to its own cluster whereas low indicate poorly match values to neighboring clusters (Sinaga and Young, 2020).

Finite cluster prototypes with their own objective functions can represent the various partitional methods. Furthermore the dissimilarity (or distance) between a point and a cluster prototype is crucial for the partition methods (Jain and Dubes 1988, MacQueen 1967). The method of k-means clustering with various extensions is very popular in the literature with application in various scientific areas (Alhawarat and Hegazi 2018).

The next step is to separate our 70 industries into distinct clusters according to their similarity or, what is the same, their homogeneity. The similarity or dissimilarity of the industries depends on the question asked and the type of industries and in our

case, the relevant question is the ranking of our industries in order of importance; namely, not all industries impart or incur the internally or externally generated shocks in the same way. Some industries are tightly connected to each other but weakly to others and others are only lightly connected with others. In our case, we categorized the 70 industries into three clusters according to the popular k-means criterion. The latter is a method of partitioning *n* observations into *k* clusters in which each industry is assigned to a particular cluster according to the nearest mean or cluster centroid around which industries are crowded. The property of *k*-means clustering is that it minimizes within-cluster variances or Euclidean distances. The mean optimizes squared errors, whereas the geometric mean minimizes Euclidean distances. Cluster analysis starts by selecting a distance measure and optimization process which meaningfully determines the number of *k* partitions or clusters (only a few) and the industries contained in each. In Figure 3 below we determine the optimal number of clusters following the Silhouette method for each of our four years of the analysis. The results show that in every particular year the optimal number of clusters is three.

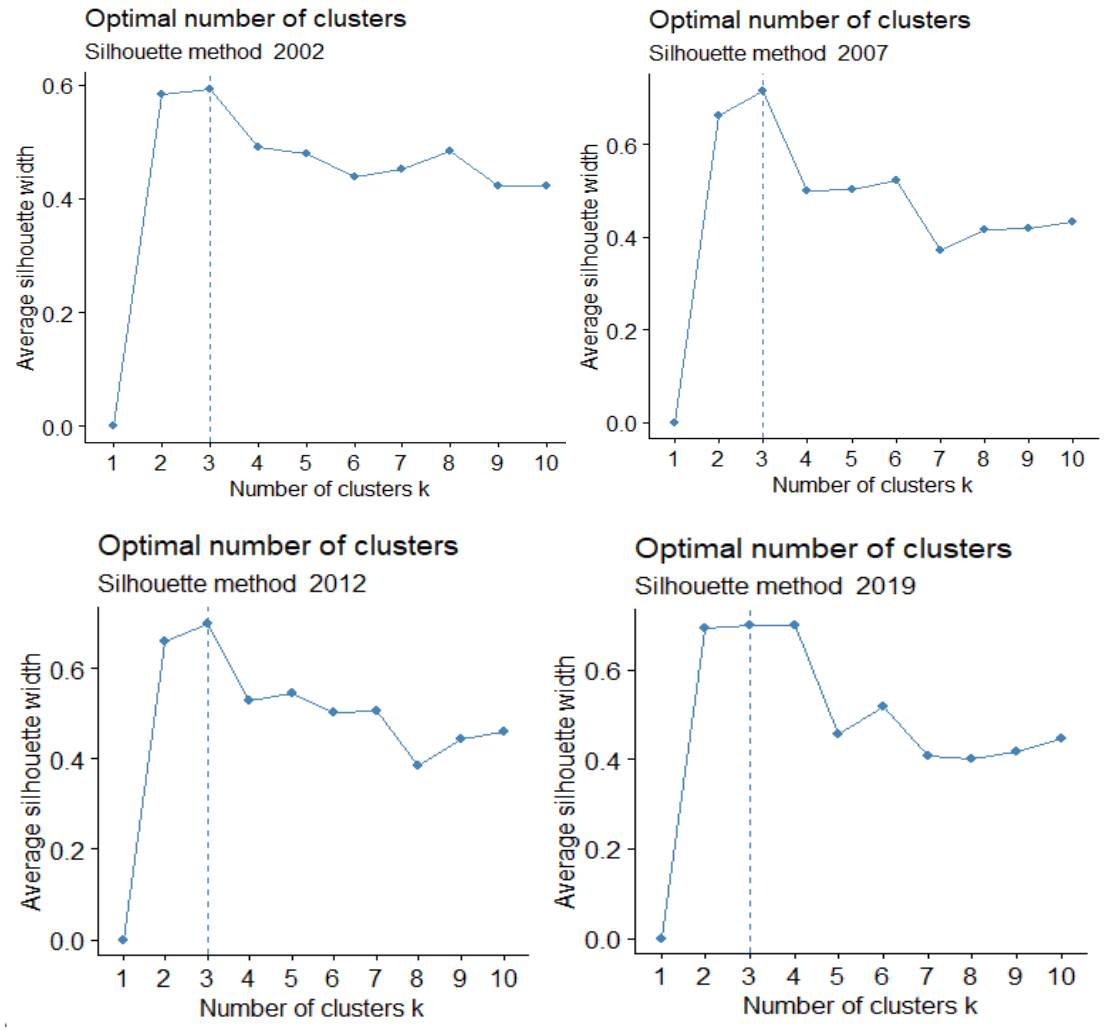


Figure 3 Optimal number clusters with silhouette method

Having established that the number of clusters is three, the industries are classified in each of these three clusters according to their respective centroid. The scree plot helps us to choose the principal components and understand the basic data structure. In Figures 4, 5, 6 and 7 we display the three distinct clusters and the number of industries contained in each.

In the interest of brevity and clarity of presentation we explain each of the four years the clusters we form as well as the industries they include. As we have already mentioned the first PC is on the horizontal axis and the second PC on the vertical axis. From the three clusters, we separate the blue or the South-East (S-E) cluster and the green or the North-East (N-E) cluster as the most important ones containing the key industries. The ranking in each particular cluster is according to the first PC and also take into account the second PC. The majority of industries are compressed in the orange or Western (W) cluster. The nomenclature of industries is in the Appendix.[4] Thus, for the benchmark input-output data of the year 2002, we have:

[4] The clusters were determined using the R programming language and also the Matlab gave the same results.

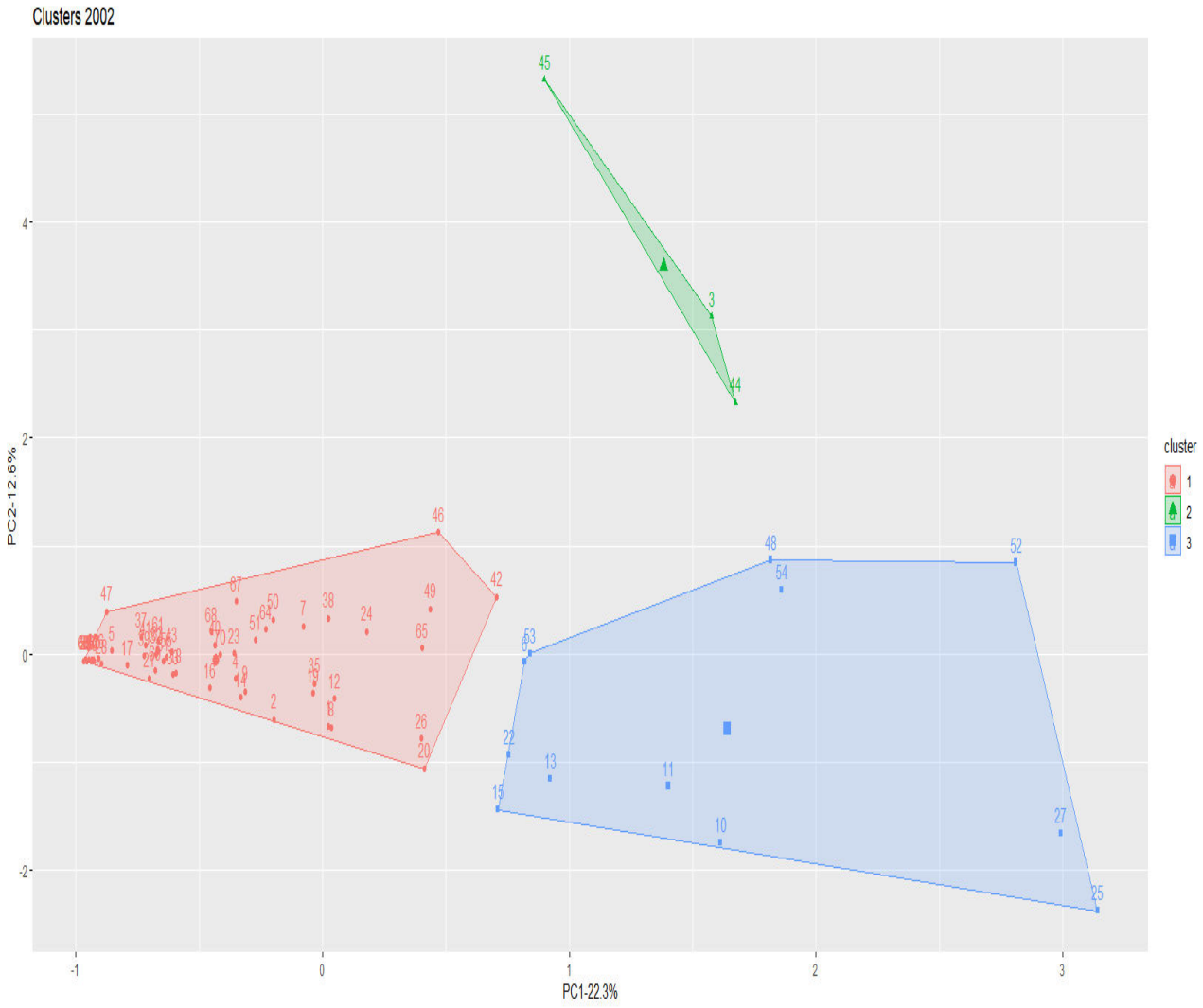


Figure 4. Clusters of industries, 2002

In Figure 4 above, the industries in the S-E (blue) and N-E (green) clusters are ranked starting from the South-East (S-E) and placing industries from the right going to the left until we exhaust the industries in the S-E cluster and then continue with the N-E cluster. Thus, we have the following ranking noting first the number of the industry and then its name. Thus, we have 25 Chemical products, 27 Wholesale trade, 52 Miscellaneous professional scientific, 54 Administrative and support services, 48 Other real estate, 10 Primary metals, 11 Fabricated metal products, 13 Computer and electronic products, 53 Management of companies and enterprises, 6 Utilities, 22 Paper products, 15 Motor vehicles bodies & trailers. Thirteen industries in total are included in the first cluster while the second paint in green cluster N-E cluster contains the following three industries: 44 Federal credit intermediation, 3 Oil and gas extraction, 45 Securities commodity contracts.[5]

[5] The nomenclature of industries is in the Appendix.

Continuing with the benchmark year 2007, we distinguish the following three clusters shown in Figure 5 below:

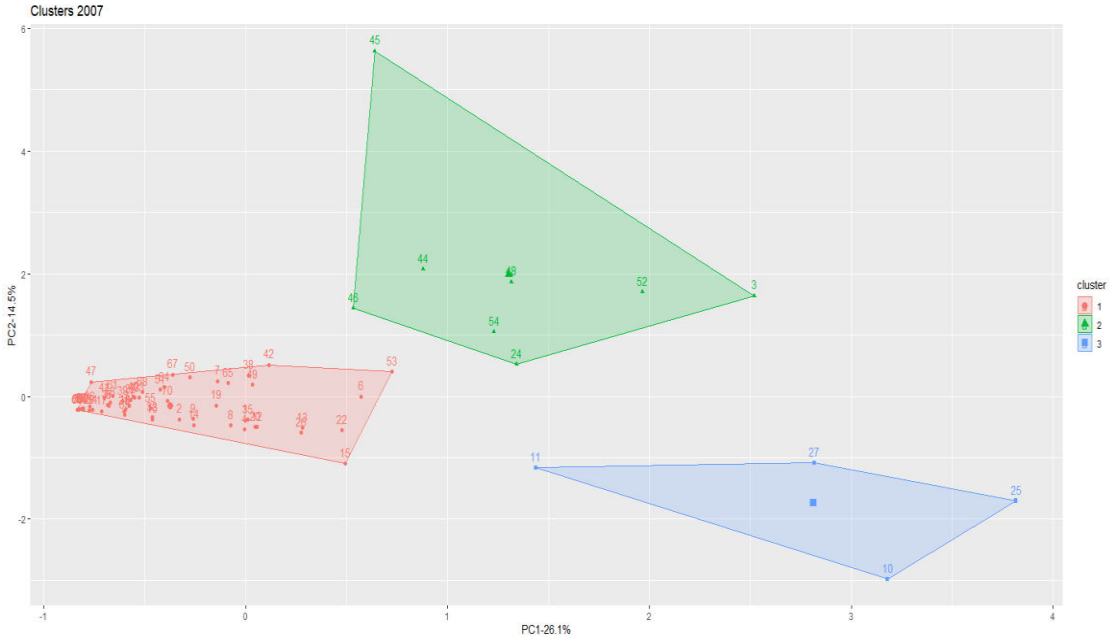


Figure 5. Clusters of industries, 2007

We observe that the number of industries in the blue or S-E cluster decreased and increased in the about North-East (green) cluster. More specifically, in dissenting order and starting from the right end of the first PC, we have the following four industries: 25 Chemical products, 10 Primary metals, 27 Wholesale trade, 11 Fabricated metal products. While the NE cluster includes more industries, which are also ranked in dissenting order, and these are: 3 Oil and gas extraction, 52 Miscellaneous profession-al scientific, 24 Petroleum and coal products, 48 Other real estate, 54 Administrative and support services, 44 Federal credit intermediation, 45 Securities commodity contracts, 46 Insurance carriers and related activities.

Continuing with the benchmark year 2012 the three clusters along with the industries contained in each are displayed in Figure 6 below:

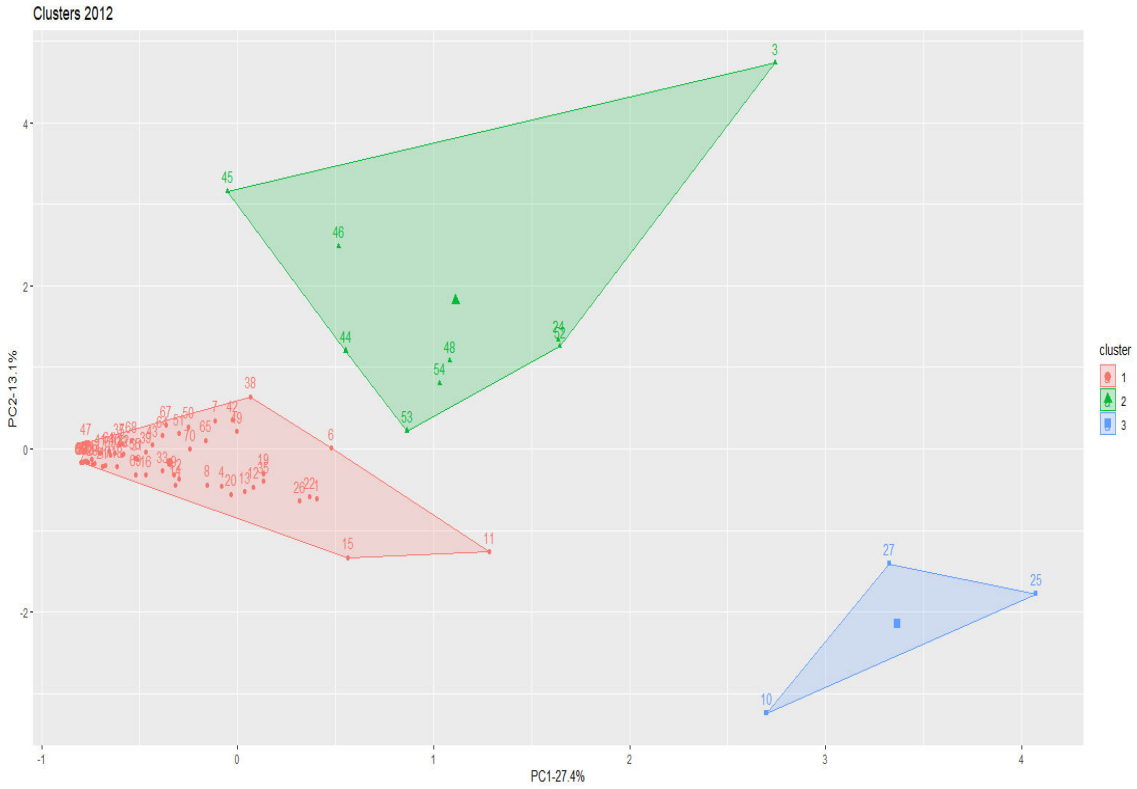


Figure 6 Clusters of industries, 2012

We observe that in the year 2012 the number of industries in the top South-East cluster dropped to only three and these are: 25 Chemical products, 27 Wholesale trade, 10 Primary metals.

The North-East cluster includes the following industries ranked in dissenting order: 3 Oil and gas extraction, 52 Miscellaneous professional scientific, 24 Petroleum and coal products, 48 Other real estate, 54 Administrative and support services, 53 Management of companies and enterprises, 44 Federal credit intermediation, 46 Insurance carriers and related activities, 45 Securities, commodity contracts etc.

Finally, the lack of a benchmark input-output table for the year 2017, which is still in its making, we utilize the last available input-output table is for the year 2019 and the three clusters are shown in Figure 7 below.

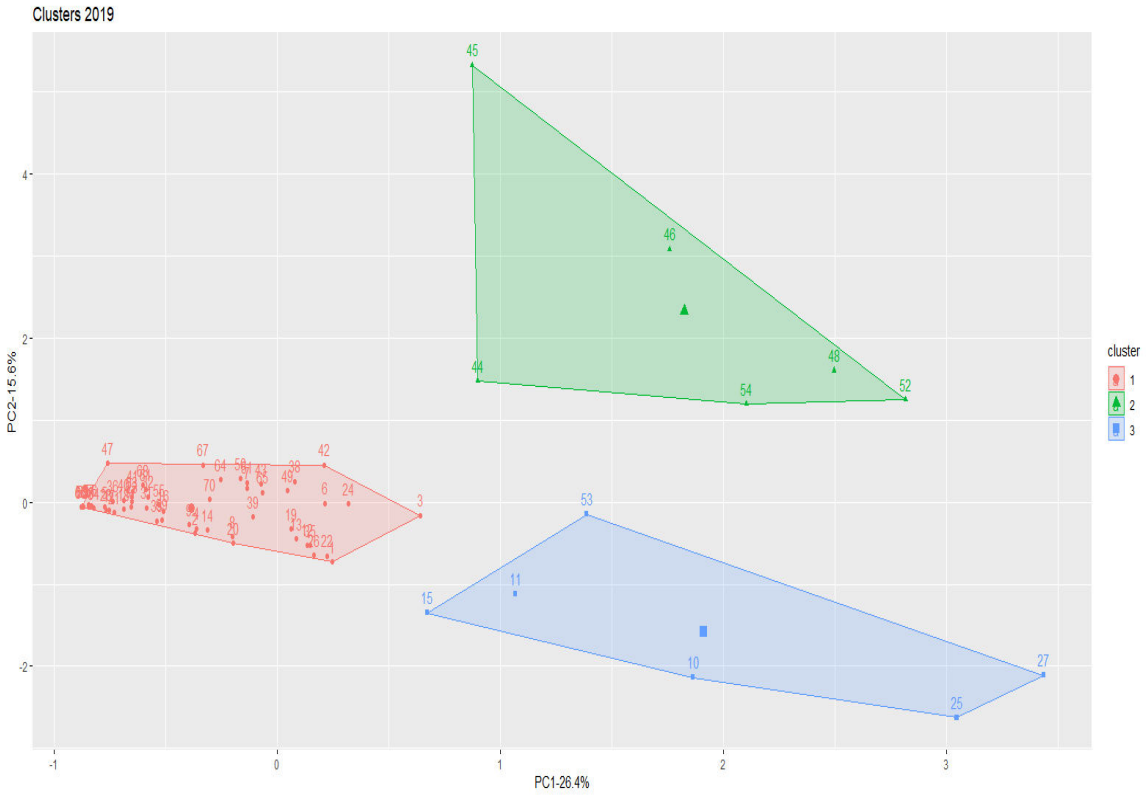


Figure 7. Clusters of industries, 2019

The blue or S-E cluster of the year 2019 is augmented to include besides the three top industries of 2012 (27 Wholesale trade, 25 Chemical products, 10 Primary metals), which are to the right end of the cluster and so continue to exert most of their influence on the economy and three more industries are added; namely, 53 Management of companies and enterprises, 11 Fabricated metal products, 15 Motor vehicles, bodies & trailers.

By contrast the green or N-E cluster contains the following industries: 52 Miscellaneous professional, scientific, 48 Other real estate, 54 Administrative and support services, 46 Insurance carriers and related activities, 45 Securities, commodity contracts and 44 Federal credit intermediation.

3. Discussion

3.1 Ranking of Industries

The clustering of industries into three groups alerts us into an altogether different vantage point of view. Hence, the grouping of industries into particular clusters ranked in order of importance makes possible the estimation of the impact of internally or externally generated shocks in the totality of the economy. For this purpose, we compare our findings from the PCA with those of total BL and FL. In effect, we experimented with the traditional techniques of identifying key industries according to the forward or BL. The results showed that neither the forward, nor the

BL, taken individually, accord to the ranking of industries found through the clustering procedure. However, by taking the total linkages, that is, the average of an industry's forward and BL, the resulting ranking is pretty close to that of the PCA. Thereby, lending support to our clustering technique as a meaningful and, at the same time, effective way of ranking industries.

The next task is to group industries into particular clusters according to how homogenous they are. The common property characterizing industries is their variance which can be classified into high, medium, and low. This clustering would make a fine example of "unsupervised learning" since we are not making predictions; we merely categorize the industries into particular groups. There is no doubt that the PCA captures better than any other parsimonious method the variance associated with each of the industries. Furthermore, the PCA enables the classification of industries according to their positive distance from zero. In particular, the further away from zero, the higher, the variance of the industry from the rest, and other similarly situated industries can be grouped forming a cluster. The location of the industry into a particular cluster indicates its association with respect to the two PC. Thus in our case, the industries grouped in the S-E cluster are characterized by high variance and therefore have a much larger effect on the economy. The second in importance N-E area, although it contains influential industries, nevertheless, they are less so, than those in the S-E. The majority of industries are crowded near zero and they are far to be considered key industries.

It is interesting to note that the ranking through the use of the average of both BL and FL is quite close to that of the PCA. The underlying idea behind the total linkages is that if higher than one, it follows that the importance of this particular industry exceeds that of the economy-wide average. The converse is in case that this figure is lower than one. We picked 1.18 as our threshold for total linkages. We found that about fifteen are the key industries, in the four years of our study, which are no different than those derived by our two clusters. Of course, we have differences in rankings, but the clustering method based on PCA gives consistent results from one benchmark year to the next and the final 2019 year. The PCA results are close to those derived using the traditional total linkages. Furthermore, the PCA possesses some additional properties that may give rise to a new research agenda and methodology in identifying clusters of industries and their possible economic impact.

An inspection of the three clusters reveals that the outer right cluster is the one whose relative importance makes it the principal cluster. Thus, the industries in the S-E, other things equal, are more influential than similarly situated industries located in the N-E cluster. It is interesting to note that industries and their ranking according to the PCA are displayed in Table 1 below. More specifically and for each of our four years, the first column displays the ranking of industries. The second column shows the industry number placed according to the traditional method. In the third column, we show the industry number contained in each cluster starting from the S-E and

continuing to N-E, and going to the W cluster until the 15th industry. There is variability between industries but not much.

Table 1. Ranking of industries through linkages and PCA clusters

2002 Rank	BL & FL (Industry)	PC (industry)	2007 Rank	BL & FL (Industry)	PC (industry)
1	1.726 (52)	3 S-E (25)	1	1.821 (25)	3 S-E (25)
2	1.637 (48)	3 S-E (27)	2	1.742 (52)	3 S-E (10)
3	1.602 (25)	3 S-E (52)	3	1.728 (10)	3 S-E (27)
4	1.539 (27)	3 S-E (54)	4	1.722 (48)	3 S-E (11)
5	1.461(54)	3 S-E (48)	5	1.582 (27)	2 N-E (03)
6	1.415 (44)	3 S-E (10)	6	1.483 (03)	2 N-E (52)
7	1.399 (10)	3 S-E (11)	7	1.438 (24)	2 N-E (24)
8	1.338 (11)	3 S-E (13)	8	1.418 (54)	2 N-E (48)
9	1.307 (15)	3 S-E (53)	9	1.402 (11)	2 N-E (54)
10	1.261 (42)	3 S-E (06)	10	1.343 (44)	2 N-E (44)
11	1.256 (03)	3 S-E (22)	11	1.305 (45)	2 N-E (45)
12	1.253 (22)	3 S-E (15)	12	1.282 (15)	2 N-E (46)
13	1.252 (13)	2 N-E (44)	13	1.260 (46)	1 W (53)
14	1.186 (45)	2 N-E (03)	14	1.234 (53)	1 W (06)
15	1.182 (46)	2 N-E (45)	15	1.224 (22)	1 W (15)

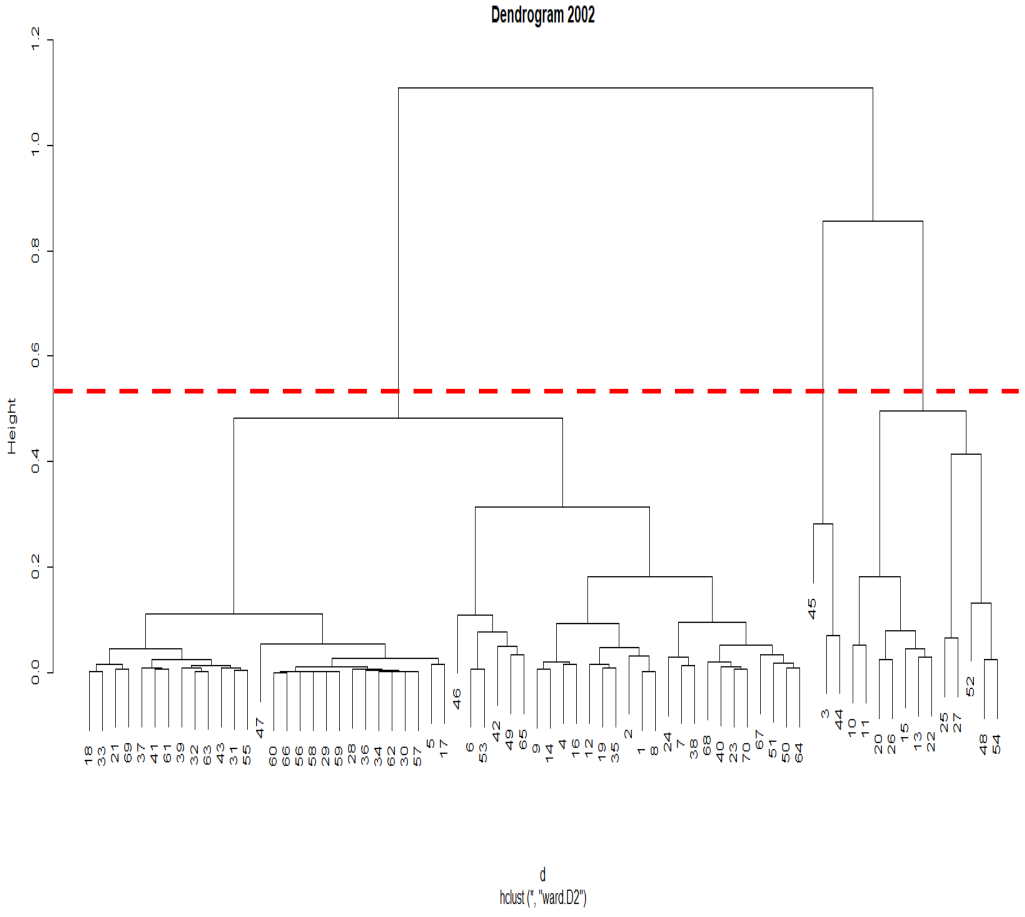
2012 Rank	BL & FL (Industry)	PCA (industry)	2019 Rank	BL & FL (Industry)	PCA (industry)
1	1.885 (25)	3 S-E (25)	1	1.855 (48)	3 S-E (27)
2	1.767 (27)	3 S-E (27)	2	1.819 (52)	3 S-E (25)
3	1.703 (10)	3 S-E (10)	3	1.770 (27)	3 S-E (10)
4	1.684 (52)	2 NE (03)	4	1.651 (25)	3 S-E (53)
5	1.613 (48)	2 NE (24)	5	1.644 (54)	3 S-E (11)
6	1.611 (24)	2 NE (52)	6	1.518 (46)	3 S-E (15)
7	1.543 (03)	2 NE (48)	7	1.514 (10)	2 NE (52)
8	1.411 (54)	2 NE (54)	8	1.336 (52)	2 NE (48)
9	1.388 (11)	2 NE (53)	9	1.336 (15)	2 NE (54)
10	1.357 (46)	2 NE (44)	10	1.320 (11)	2 NE (46)
11	1.323 (15)	2 NE (46)	11	1.258 (24)	2 NE (44)
12	1.245 (19)	2 NE (45)	12	1.218 (45)	2 NE (45)
13	1.241 (53)	1 W (11)	13	1.213 (03)	1 W (03)
14	1.211 (22)	1 W (15)	14	1.208 (44)	1 W (24)
15	1.201 (44)	1 W (06)	15	1.199 (19)	1 W (01)

3.1 Industries Ordered in Dendrograms

A salient feature of the PCA is that clustering enables the presentation of the industries in a dendrogram, which marks the last step in cluster analysis. A

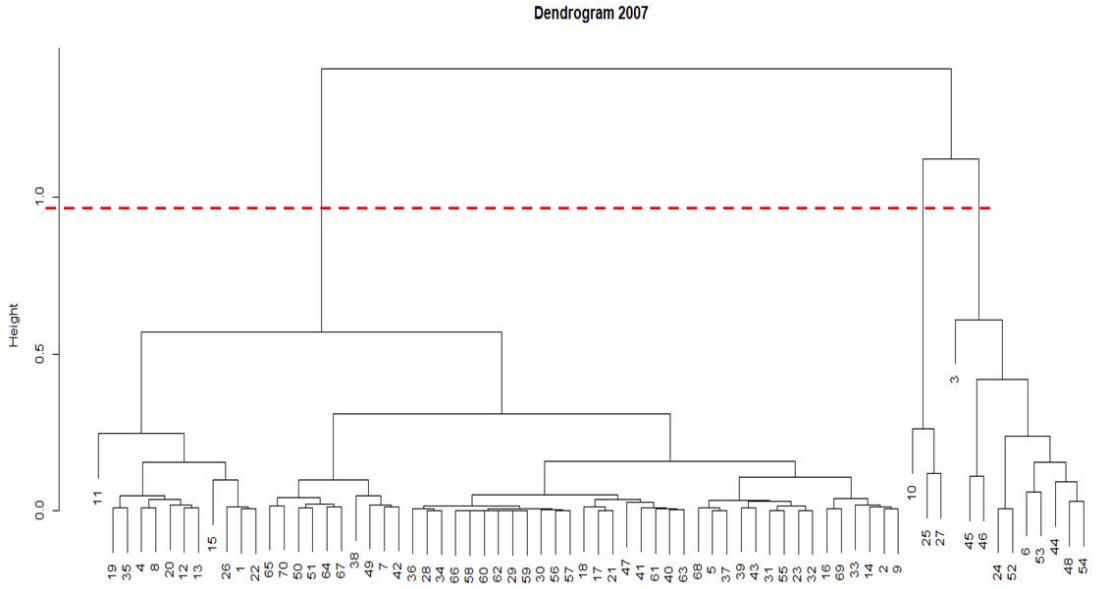
dendrogram is a hierarchical tree plot that displays a grouping of industries into distinct clusters. The length of each branch on the graph measures the distance between industries in the cluster. The purpose of dendrograms is to decide upon the suitable number of clusters. For this purpose, we employ the agglomerative method, which creates a hierarchy of industries starting with all of them as if they were completely separated and then fuses them until there is only one cluster left.

The dendrograms below mark the presence of three clusters. The clusters are distinguished starting from the horizontal axis, where all 70 industries stand like the leaves of the tree. As we move upward, ideally drawing a parallel line to the hypothetical horizontal axis, we start distinguishing the branches directed to the core of the imaginary tree. In this process, we end up with three main branches for each year of our study, exactly as indicated by our k-means testing procedure. In each of these branches, we identify not only the tree clusters but also their detailed connections. We observe a hierarchical location of the industries ranked according to the height of the branches. The higher a branch, the higher the relative importance of the industry.

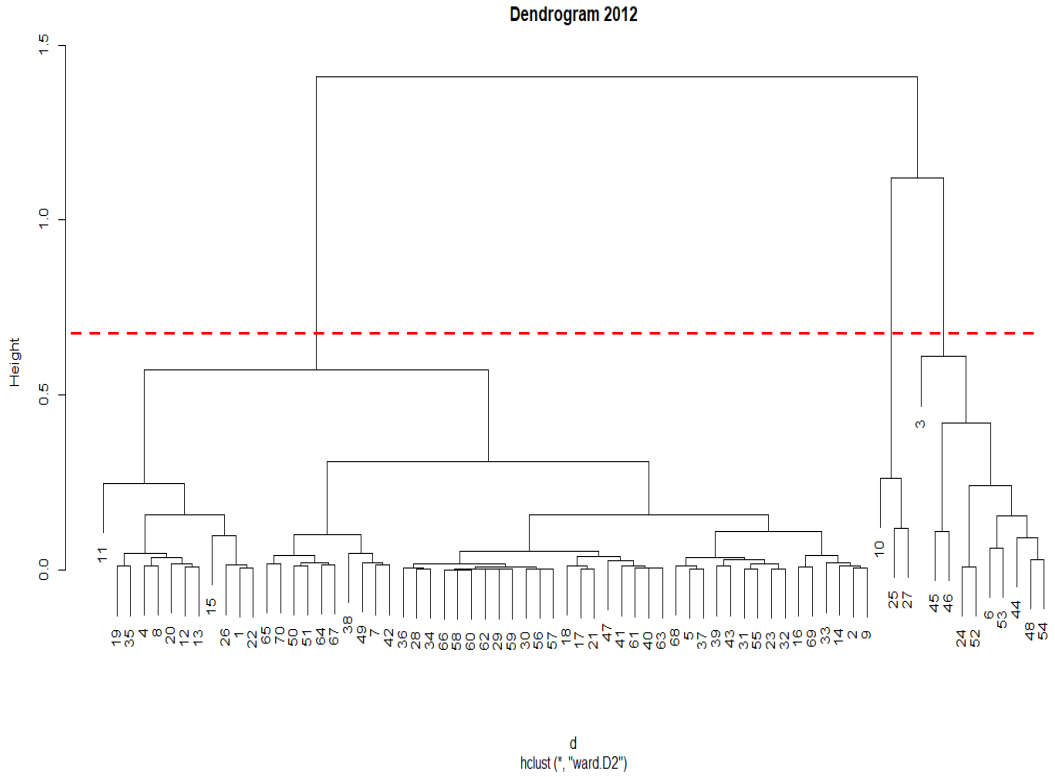


The nodes that are higher indicate the importance of each cluster and within the same cluster the importance of the industry. In other words the longer the branch the more

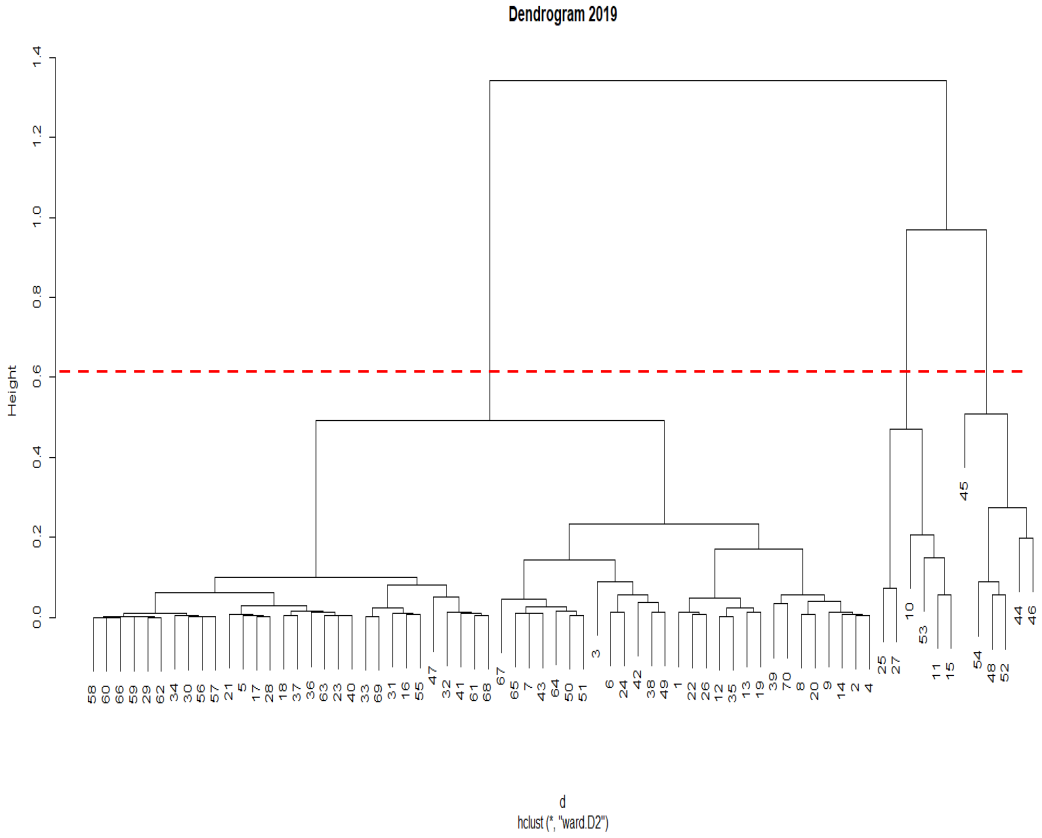
important the industries that branch out of it. And from each branch the industries located above the others carry more weight than those being below it. Thus starting from the benchmark year 2002 the dashed red line indicates the presence of three clusters and the top one consists in dissenting order of industries (45, 3 and 44), which make up the first main brunch. Going to the next in importance branch the industries 52, 48, 54 industries 25, 27 follow 15, 13, 22 and so forth.



In the dendrogram of the year 2007, we have the following ranking in descending order 10, 25, 27 from the first major brunch and for the second 3, 45, 46, 24, 25, 44, 6, 53, 48 and 54.



The same ranking with respect to the top longer brunch is repeated in the dendrogram of the year 2012. Thus, we have industries 10, 25, 27 followed by 3, 45, 46, 24, 25, 6, 53, 44, 48 and 54.

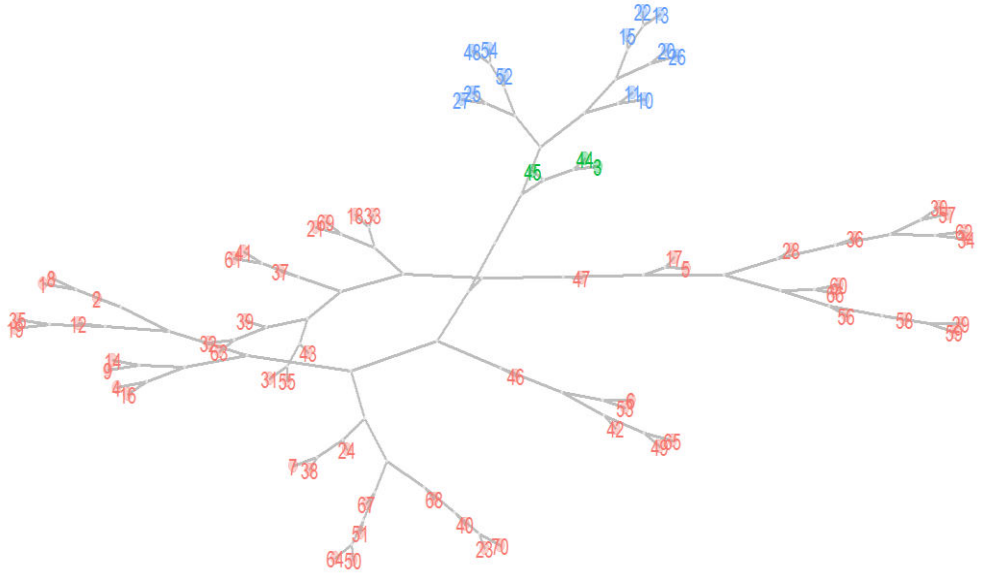


Finally, the dendrogram of the year 2019 gives the following ranking of industries 25, 27, 10, 53, 11, 15, 45, 44, 46, 54, 48, 52

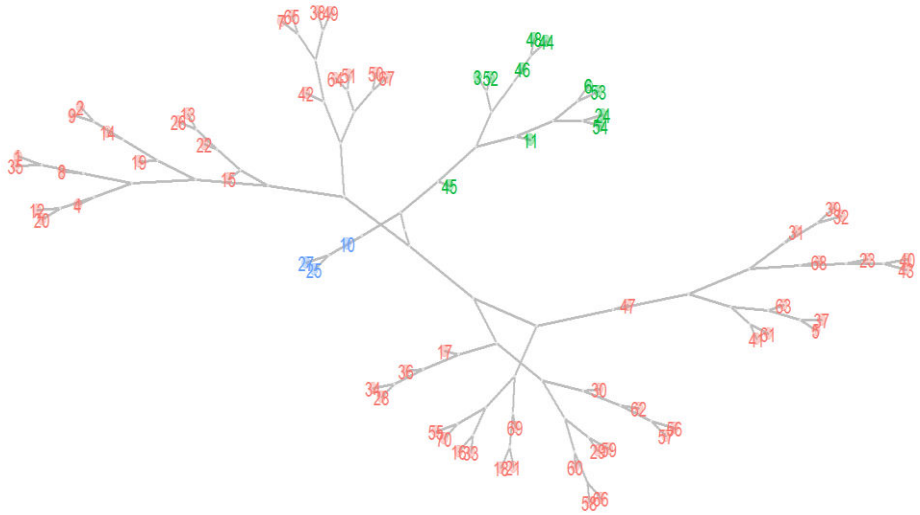
Dendrograms may be proved particularly helpful in our understanding of the formation and the internal structure of clusters, and they can be profitably used in laying bare underlying trends and highlighting outliers. Such information is unquestionably practical in tracking down the process of structural change and technological change. For this purpose, the more informed inter-cluster and intra-cluster connections of industries shed more light on all of the above. The panel of four dendrograms displayed in Figure 5 provides us with a visual description of such inter and intracluster connections of industries.

A similar picture is obtained by looking at the particular networks formed by industries displayed in a panel of four graphs in Figure 5 below, where the clusters are painted in green, blue, and red colors to be distinguished from each other

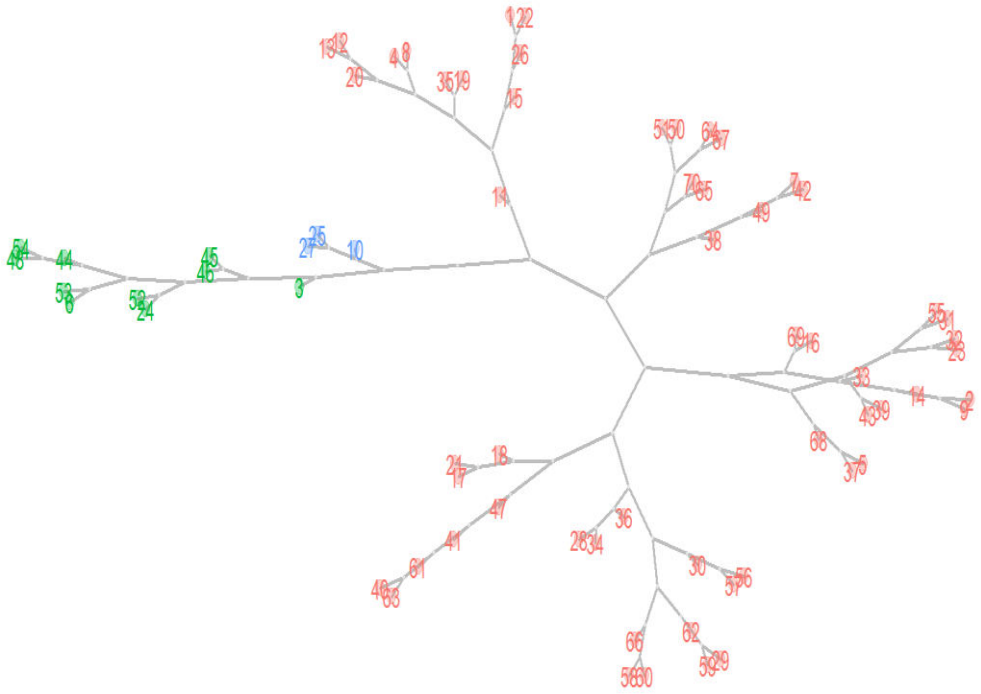
2002 network



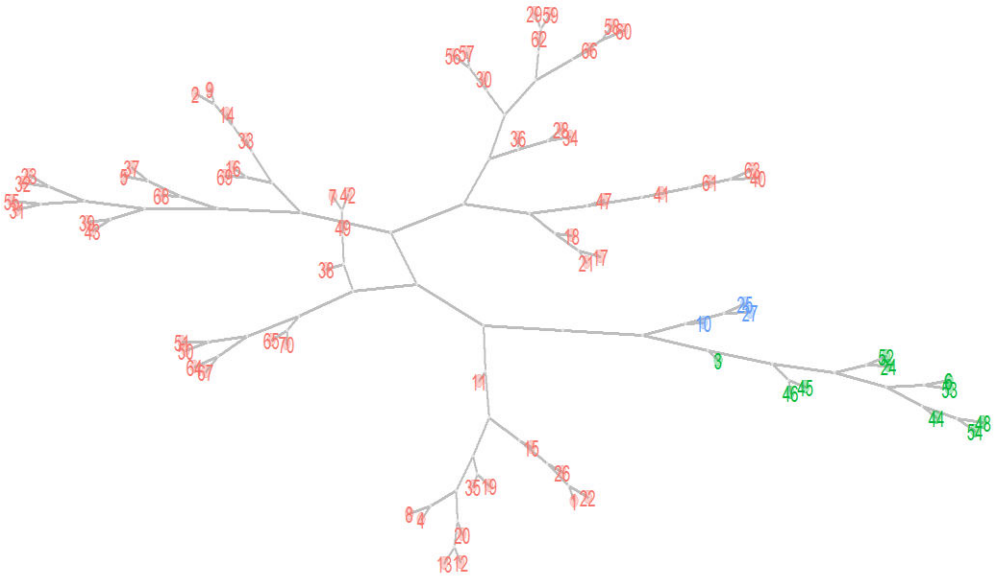
2007 network



2012 network



2019 network



4. Concluding Remarks

In evaluating the key industries grouped into clusters and basic economic structures, the PCA possesses distinct advantages compared to the standard methods. The PCA, a mathematically rigorous and parsimonious technique, enables the more efficient utilization of input-output data. As a result, the PCA is not limited merely to the ranking of industries, as in the standard methods, but it further refines the ranking of industries by ordering them into well-defined clusters. In so doing, the PCA expands the identification of key industries in new directions. Thus, starting with clusters and going into dendrograms and networks, we identify the connections between clusters and the industries within clusters.

In experimenting with the data from our four input-output tables of the US economy, we observed a close association between the leading PCs and the total linkages of industries. A result that encourages the use of the PCA and its application to input-output data for the identification of key industries. The next step was to use the top two PCs perpendicular to each other, meaning their correlation is zero. We have also utilized these two PCs as the horizontal and vertical axis, respectively. In such representation, we grouped the data into three particular clusters for each of the four distant years of our study. The clustering of industries was decided through the use of the k-means and silhouette procedures. It is interesting to note that approximately the same industries in the top two clusters are repeated over the years and the very few that are not included stand as border cases. Furthermore, the industries in the top two clusters are no different from those derived from the average of the backward and forward linkages. The salient feature of the PCA is the grouping of industries into clusters and their arrangement into dendrograms. The network presentation reveals the interlinkages of industries within clusters as well as the hierarchical positions of clusters and their connections.

Availability of data and materials

The datasets for the present study are available electronically and publicly. Data on input–output tables are available from the website of the Bureau of Economic Analysis.

Declarations

Ethics approval and consent to participate

Not Applicable

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Funding

The authors have not received funding from any sources for this research.

Authors' contributions

Not applicable.

Acknowledgments

Not Applicable

Authors' information

Lefteris Tsoulfidis, Professor

Department of Economics, University of Macedonia, Thessaloniki, Greece, Lnt@uom.edu.gr

Ioannis Athanasiadis, Doctorate Candidate

Department of Economics, University of Macedonia, Thessaloniki, Greece, athang@uom.edu.gr

Corresponding author

Correspondence to Lefteris Tsoulfidis: Lnt@uom.edu.gr

References

Alhawarat M. & Hegazi M. (2018). Revisiting K-Means and topic modeling, a comparison study to cluster Arabic documents, *IEEE Access*, vol. 6, pp. 42740-42749.

Bozdogan H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions, *Psychometrika*, vol. 52, no. 3, pp. 345-370.

Calinski T. & Harabasz J. (1974). A dendrite method for cluster analysis, *Commun. Statist. Theory Methods*, vol. 3, no. 1, pp. 1-27.

Davies D. L. & Bouldin D. W. (1979). A cluster separation measure, *IEEE Trans. Pattern Anal. Mach. Intell.*, vols. PAMI-1, no. 2, pp. 224-227.

Dempster A. P., Laird N. M., & Rubin D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion), *J. Roy. Stat. Soc*., Ser. B, Methodol., vol. 39, no. 1, pp. 1-38,.

Dunn J. C. (1973). A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters, *J. Cybern*., vol. 3, no. 3, pp. 32-57.

Farné M. & Vouldis A. (2021). Banks' business models in the euro area: a cluster analysis in high dimensions. *Annals of Operations Research* https://doi.org/10.1007/s10479- 021-04045-

Halkidi M., Batistakis Y., & Vazirgiannis M.(2001). On clustering validation techniques, *J. Intell. Inf. Syst.*, vol. 17, no. 2, pp. 107-145.

Iliadi F., Mariolis T., Soklis G. & Tsoulfidis L. (2014). Bienenfeld's approximation of production prices and eigenvalue distribution: further evidence from five European economies, *Contributions to Political Economy*, 33(1), 35-54.

Jain A. K. (2010). Data clustering: 50 years beyond K-means, *Pattern Recognit. Lett*., vol. 31, no. 8, pp. 651-666.

Jain A.K. & Dubes R.C. (1988). Algorithms for Clustering Data, Englewood Cliffs, NJ, USA: Prentice-Hall.

Kass R. E. & Raftery A. E.(1995), Bayes factors, *J. Amer. Stat. Assoc*., vol. 90, pp. 773-795.

Kaufman L. & Rousseeuw P.J. (1990). Finding Groups in Data: An Introduction to Cluster Analysis. New York, NY, USA: Wiley.

Lv Z., Liu T., Shi C., Benediktsson J.,& Du H. (2019). Novel land cover change detection method based on k-Means clustering and adaptive majority voting using bitemporal remote sensing images, *IEEE Access*, vol. 7, pp. 34425-34437.

MacQueen J. (1967). Some methods for classification and analysis of multivariate observations, in *Proc. 5th Berkeley Symp. Math. Statist. Probab*., vol. 1, pp. 281-297.

Mariolis T., & Tsoulfidis L. (2018). Less is more: Capital theory and almost irregular uncontrollable actual economies. *Contributions to Political Economy*, 37(1), 65–88.

McLachlan G. J. & Basford K. E. (1988). Mixture Models: Inference and Applications to Clustering. New York, NY, USA: Marcel Dekker.

Meng Y., Liang J., Cao F., & He Y. (2018). A new distance with derivative information for functional k-means clustering algorithm, *Inf. Sci*., vols. 463-464, pp. 166-185.

Meyer C. (2001). Matrix Analysis and Applied Linear Algebra. *New York: Society for Industrial and Applied Mathematics.*

Miller R. & Blair P. (2009). Input–Output Analysis: Foundations and Extensions. New York: Cambridge University Press.

Plerou V., Gopikrishnan P., Rosenow B., Amaral L., Guhr T. & Stanley E. (2002). Random matrix approach to cross correlations in financial data, *Physics Review E* 65(6), 1-18.

Rousseeuw P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, *J. Comput. Appl. Math*., vol. 20, pp. 53-65.

Sinaga K. P. & Yang, M. S. (2020). Unsupervised K-means clustering algorithm, *IEEE Access*, 8, 80716-80727.

Tsoulfidis L. (2021). Capital Theory and Political Economy: Prices, Income Distribution and Stability. London: Routledge.

Yang M.-S., Chang-Chien S.-J., & Nataliani Y. (2018), A fully-unsupervised possibilistic C-Means clustering algorithm, *IEEE Access*, vol. 6, pp. 78308-78320.

Yu J., Chaomurilige C., & Yang M.-S. (2018), On convergence and parameter selection of the EM and DA-EM algorithms for Gaussian mixtures, *Pattern Recognit*., vol. 77, pp. 188-203.

Zhu J., Jiang Z., Evangelidis G. D., Zhang C., Pang S., & Li Z. (2019), Efficient registration of multi-view point sets by K-means clustering, *Inf. Sci*.,vol. 488, pp. 205-218.

Appendix. Nomenclature of industries

No	Industries	No	Industries

1	Farms
2	Forestry, fishing, and related activities
3	Oil and gas extraction
4	Mining, except oil and gas
5	Support activities for mining
6	Utilities
7	Construction
8	Wood products
9	Non-metallic mineral products
10	Primary metals
11	Fabricated metal products
12	Machinery
13	Computer and electronic products
14	Electrical equipment appliances, etc.
15	Motor vehicles, bodies & trailers
16	Other transportation equipment
17	Furniture and related products
18	Miscellaneous manufacturing
19	Food, beverage and tobacco
20	Textile mills and textile product mills
21	Apparel and leather and allied products
22	Paper products
23	Printing and related support activities
24	Petroleum and coal products
25	Chemical products
26	Plastics and rubber products
27	Wholesale trade
28	Motor vehicle and parts dealers
29	Food and beverage stores
30	General merchandise stores
31	Other retail
32	Air transportation
33	Rail transportation
34	Water transportation
35	Truck transportation
36	Transit and ground pass. transportation
37	Pipeline transportation
38	Other transport. and support activities
39	Warehousing and storage
40	Publishing, except internet
41	Motion picture and recording industries
42	Broadcasting and telecommunications
43	Data processing, internet publishing, etc.
44	Fed., credit intermediation, etc.
45	Securities, commodity contracts, etc.
46	Insurance carriers and related activities
47	Funds, trusts, and other financial vehicles
48	Other real estate
49	Rental and leasing services etc.
50	Legal services
51	Computer systems design etc.
52	Miscellaneous professional, scientific, etc.
53	Management of companies and enterprises
54	Administrative and support services
55	Waste management and remediation services
56	Educational services
57	Ambulatory health care services
58	Hospitals
59	Nursing and residential care facilities
60	Social assistance
61	Perform. arts, spectator sports, museums
62	Amusements, gambling, and recreation
63	Accommodation
64	Food services and drinking places
65	Other services, except government
66	Federal general government (defense)
67	Federal general government (nondefense)
68	Federal government enterprises
69	State and local general government
70	State and local government enterprises