**UNIVERSIDAD PONTIFICIA COMILLAS**
ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA (ICAI)
GRADO EN INGENIERÍA EN TECNOLOGÍAS INDUSTRIALES

GRADO EN INGENIERÍA EN TECNOLOGÍAS INDUSTRIALES

TRABAJO DE FIN DE GRADO

# Using PCA and K-means clustering in input/output matrices to understand structural evolution in economies

Autor: Eloy Ramón de Sola Lantero

Director: Pablo Dueñas Martínez

Co-Director: Miguel Vázquez Martínez

Madrid

**UNIVERSIDAD PONTIFICIA COMILLAS**
ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA (ICAI)
GRADO EN INGENIERÍA EN TECNOLOGÍAS INDUSTRIALES

Declaro, bajo mi responsabilidad, que el Proyecto presentado con el título

**"USING PCA AND K-MEANS CLUSTERING IN INPUT/OUTPUT MATRICES TO UNDERSTAND STRUCTURAL EVOLUTION IN ECONOMIES"**

en la ETS de Ingeniería - ICAI de la Universidad Pontificia Comillas en el curso académico 2023/2024 es de mi autoría, original e inédito y no ha sido presentado con anterioridad a otros efectos.

El Proyecto no es plagio de otro, ni total ni parcialmente y la información que ha sido tomada de otros documentos está debidamente referenciada.

Fdo.: Eloy Ramón De Sola Lantero Fecha: 30/07/ 2024

Autorizada la entrega del proyecto

EL DIRECTOR DEL PROYECTO

Fdo.: Pablo Dueñas Martínez Fecha: 30/07/2024

Autorizada la entrega del proyecto

EL CO-DIRECTOR DEL PROYECTO

Fdo.: Miguel Vázquez Martínez Fecha: 30/07/2024

GRADO EN INGENIERÍA EN TECNOLOGÍAS INDUSTRIALES

TRABAJO DE FIN DE GRADO

# Using PCA and K-means clustering in input/output matrices to understand structural evolution in economies

Autor: Eloy Ramón de Sola Lantero

Director: Pablo Dueñas Martínez

Co-Director: Miguel Vázquez Martínez

Madrid

**UNIVERSIDAD PONTIFICIA COMILLAS**
Escuela Técnica Superior de Ingeniería (ICAI)
Grado en Ingeniería en Tecnologías Industriales

# USING PCA AND K-MEANS CLUSTERING IN INPUT/OUTPUT MATRICES TO UNDERSTAND STRUCTURAL EVOLUTION IN ECONOMIES

Author: De Sola Lantero, Eloy Ramón.

Supervisor: Dueñas Martínez, Pablo.

Co-Supvervisor: Vazquez Martínez, Miguel

Collaborating Entity: ICAI – Universidad Pontificia Comillas

## ABSTRACT

**Keywords**: Leontief Matrix, Machine Learning, Principal Component Analysis, Clustering.

This project focuses on generating and analysing the Leontief matrix using machine learning techniques implemented through MATLAB. Combining these methods to enhance the understanding of the interdependencies in the economy for them to be considered in future policy making. the objective is to generate an improvement in predictive capabilities and accuracy compared with traditional methods.

Objectives

- Develop a machine learning model for analysing the Leontief Matrix.
- Implement the model with MATLAB.
- Evaluate the predictive accuracy of the model compared to traditional methods.

The Leontief Input-Output matrix is a fundamental tool in economics for understanding the interdependencies in the economy for changes of output caused by changes in demand. It is constructed from an input-output matrix in which rows represent the output product of each industry, columns represent those same industries but as inputs to generate their output. This resource has been studied and varied to analyse several areas of the economy, such as the changes in output caused by variance in employment and their income.

Machine learning techniques such as Principal Component Analysis (PCA) and K-means clustering can significantly improve the analysis of the Leontief matrix. PCA focuses on reducing the dimensionality of the analysis, simplifying the objects in a two-dimensional space that maximizes the variability represented. Making it easier to understand the key patterns and trends. K-means is more focused in agglomerating industries that share similar behaviours. For this study in particular it will group industries based on their relation in changes in output generation based on specific changes in demand.

## Methodology

Data will be collected from Bureau of Labor Statistics (BLS), since the input-output matrix they build up is divided into 192 and is based on the North American Industrial Classification System (NAICS) 2022. It is grouped based production processes, input requirements output characteristics, market orientation and technological similarities. The institution has been collecting data since 1997 and carries out revisions so that the data matches reality to the closest extent. We will choose 2012 data so that it has been revised and so that we can find articles and studies on the years economy functioning.

Specifically, the "USE" matrix shows the sales to intermediate consumers and a final demand. Added value on the goods produced is eliminated from the equation for a better understanding on raw output. From this matrix the Leontief model will be build, obtaining the Leontief multipliers.

Various machine learning techniques are implemented in MATLAB, including principal component analysis and clustering, to identify patterns and predict economic outcomes. Highlighting the results obtained, limitations and potential future works.

## Leontief Matrix

To obtain the Leontief Matrix the total output of each industry needs to be added up, including the final demand, which is represented by each row in the USE matrix, represented by $X_i$. Each element of the intermediate matrix ($Z_{ji}$) needs to be divided by the total output to obtain the direct element matrix A. This matrix needs to subtract the identity matrix (I) and to be inverted into the Leontief matrix.

$$A = [a_{ij}] = [\frac{Z_{ij}}{X_i}]$$

*Ecuation 1: Direct element matrix*

$$L = (I - A)^{-1}$$

*Equation 1: Leontief Matrix*

This is composed of the Leontief multipliers $l_{ij}$, which represent how the change in demand of one unit of industry j affects the output of the industry i.

## MACHINE LEARNING

For all techniques used in this category it is crucial to understand that industries output acts as objects and the changes in demand act as variables due to the configuration of the matrix.

### Principal component analysis

When carrying out the PCA on the dataset industries as observation as red dots in the two dimensional space, whereas the blue lines represent the contribution of each variable as to the principal component on each dimension.

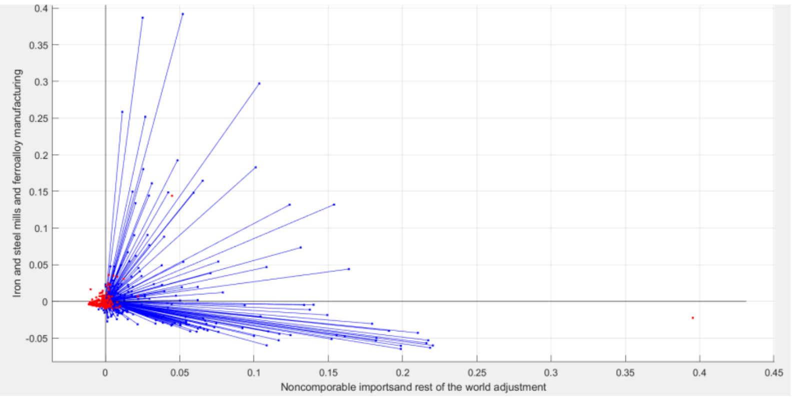


Figure 2 (repeated): PCA sample.

Through the principal component analysis carried out on the whole database it was revealed Noncomparable imports and rest of the world adjustment is the most dominant variable, explaining 10.27% of the total variance. Later it was discovered that this representation of the variance is strongly explained by the heavy dependence on the industry by the economy. Apart from this relevant result, it was difficult to discern any other information in this complex economy, this is why it was crucial to carry out the sectorial PCA.

This implies carrying out PCA on the agglomerations of industries on their own. According to the North American Industrial Classification System (NAICS) industries can be furtherly aggregated into sectors. Based on their criteria, it would be expected that industries in a same sector will be strongly interconnected. PCA has been carried out on a sector which act as objects and all industries act as variables.

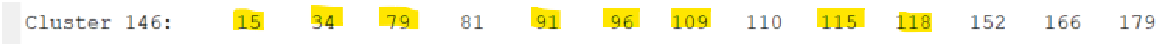
When doing this we extracted the following information. Most sectors feel little to no effect from industries outside their sector, especially the Finance and Insurance, and Health Care. The only sector that was strongly affected by changes in demand outside their sector was Special Industries, which includes the two principal components of the general PCA.

### K-means clustering

This method involves the finding of non-overlapping clusters and their centres in a set of unlabelled data such as the one we have. The decisions include the number of clusters one desires to obtain from the data and what method to use as clustering criteria.

To find out this value an Elbow Method analysis was carried out. It is based on the minimisation of Within-Cluster Sum of Squares (WCSS), which is the sum of the squared distances between each point and the centroid of its cluster. Through two methods of calculating the optimum k, the data was optimally represented by 180 clusters.

When clustering using this value of k and the Euclidean distance, we reached an inconclusive result. The composition of clusters with more than one element varied between iterations, with only a few values repeating consistently.



```
Cluster 146:    15   34   79   81   91   96   109   110   115   118   152   166   179
```

*Figure 1: Composition of composite cluster, with highlighted repetitions*

The industries highlighted were the ones to constantly appear. Some arguments may explain the relations from different points of view. Such as sector similarity between Construction (15) and Real State (118), due to the symbiotic relation they share. The relations between these industries seem to be clear on a bilateral basis, but it is extremely hard to find a common nexus between all these industries due to the complexity of the picture. These are the only relations that highlight in such a complex economy where all objects seem to not share similarities with others.

When carrying out the K-means for the lowest possible value to see which is the most different industry Noncomparable imports and rest of the world adjustment was the first to form its own cluster. Which is due to its dependence on the whole economy’s changes in demand.

## CONCLUSIONS AND FUTURE WORK

PCA highlighted the existence of some key industries that influence the output of the rest such as Noncomparable imports and some affect certain sectors such as Steel Mills and Ferroalloy Manufacturing on the Special industries. K-means analysis highlighted the complexity of the output relations, with a high number of optimal clusters indicating each industry's unique contribution to economic variability. This statement is also supported by the fact that the members of the cluster with more than one industry varied with iterations.

Due to the revealed complexity of the economy, policies should consider the interdependencies between industries in a same sector. An example would be incentivising the Insurance Carriers demand to affect the housing sector. Symbiotic relationships, like this between construction and real estate, suggest that policies affecting one could have a ripple effect on the other.

This analysis has found some limitations. As the elbow methos and silhouette score results were inconsistent due to the complexity of the data sample. PCA’s reduction to two dimensions may lead to a loss of information, especially which such a dispersed variability. The study's focus on a single year and older data limits its applicability to current economic conditions, but it will not possess actual reports on the industries behaviour.

For future research it should be considered the analysing of smaller, simpler economies to identify patterns that can be applied to the American economy. Incorporating variables such as inflation effects on industry’s output could provide more comprehensive reports. Testing predictive models could answer what interrelations found in this analysis may apply to effective policy making.

**UNIVERSIDAD PONTIFICIA COMILLAS**
ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA (ICAI)
GRADO EN INGENIERÍA EN TECNOLOGÍAS INDUSTRIALES

# USO DE PCA Y K-MEANS CLUSTERING EN MATRICES INPUT/OUTPUT PARA ENTENDER LA EVOLUCIÓN ESTRUCTURAL EN LAS ECONOMÍAS

Autor: De Sola Lantero, Eloy Ramón.

Director: Dueñas Martínez, Pablo.

Codirector: Vázquez Martínez, Miguel.

Entidad Colaboradora: ICAI – Universidad Pontificia Comillas.

## RESUMEN DEL PROYECTO

**Palabras Clave**: Matriz de Leontief, Aprendizaje Automático, Análisis de Componentes Principales, Clustering.

Este proyecto se enfoca en generar y analizar la matriz de Leontief utilizando técnicas de aprendizaje automático implementadas a través de MATLAB. Combinando estas herramientas para mejorar la comprensión de las interdependencias en la economía para que sean consideradas en el desarrollo de políticas económicas futuras. El objetivo es generar una mejora en las capacidades predictivas y precisión en comparación con los métodos tradicionales.

Objetivos

- Desarrollar un modelo de aprendizaje automático para analizar la Matriz de Leontief.
- Implementar el modelo con MATLAB.
- Evaluar la precisión predictiva del modelo en comparación con los métodos tradicionales.

La matriz de Input-Output de Leontief es una herramienta fundamental en economía para entender las interdependencias en la economía, en particular los cambios en la producción debidos a cambios en la demanda. Se construye a partir de una matriz de input-output en la que las filas representan el producto de salida de cada industria, y las columnas representan

esas mismas industrias, pero como inputs para generar su producción. Este recurso ha sido estudiado y alterado para analizar varias áreas de la economía, como los cambios en la producción causados por la variabilidad en el empleo y sus ingresos.

Las técnicas de aprendizaje automático como el Análisis de Componentes Principales (PCA) y el agrupamiento K-means pueden mejorar significativamente el análisis de la matriz de Leontief. PCA se centra en reducir la dimensionalidad del análisis, simplificando los objetos en un espacio bidimensional que maximiza la variabilidad representada, facilitando la comprensión de los patrones y tendencias clave. K-means se enfoca más en agrupar industrias que comparten comportamientos similares. Para este estudio en particular, se agrupará industrias basándose en su relación con los cambios en producción según cambios específicos de demanda.

**Metodología**

Se recopilarán datos de la Oficina de Estadísticas Laborales (BLS), ya que la matriz input-output que elaboran está dividida en 192 sectores y se basa en el Sistema de Clasificación Industrial de América del Norte (NAICS) 2022. Está agrupa según procesos de producción, requisitos de inputs, características de producto de salida, orientación al mercado y similitudes tecnológicas. La institución ha estado recopilando datos desde 1997 y realiza revisiones para que los datos coincidan con la realidad. Elegiremos datos de 2012 para que hayan sido revisados y para poder encontrar artículos y estudios sobre el funcionamiento económico de esos años.

Específicamente, la matriz "USE" muestra las ventas a consumidores intermedios y una demanda final. El valor añadido en los bienes producidos se elimina de la ecuación para una mejor comprensión de la producción bruta. A partir de esta matriz se construirá el modelo de Leontief, obteniendo sus respectivos multiplicadores.

Varias técnicas de aprendizaje automático se implementarán en MATLAB, incluyendo el análisis de componentes principales y el agrupamiento, para identificar patrones y predecir resultados económicos. Se destacarán los resultados obtenidos, las limitaciones y los posibilidades para trabajos futuros

### Matriz de Leontief

Para obtener la Matriz de Leontief, se debe sumar la producción total de cada industria, incluida la demanda final, que está representada por cada fila en la matriz USE, representada por Xi. Cada elemento de la matriz intermedia (Zji) debe dividirse por la producción total para obtener la matriz de elementos directos A. Esta matriz debe restar la matriz identidad (I) y ser invertida. Esta matriz obtenida como resultado es la matriz de Leontief.

$$A = [a_{ij}] = [\frac{Z_{ij}}{X_i}]$$

*Ecuación 2: Matriz de elementos directos*

$$L = (I - A)^{-1}$$

*Ecuación 2: Matriz de Leontief*

Esta se compone de los multiplicadores de Leontief $L_{ij}$, que representan cómo el cambio en la demanda de una unidad de la industria j afecta la producción de la industria i.

## APRENDIZAJE AUTOMÁTICO

Para todas las técnicas utilizadas en esta categoría, es crucial entender que la producción de las industrias actúa como objetos y los cambios en la demanda actúan como variables debido a la configuración de la matriz.

### Análisis de componentes principales

Al realizar el PCA en el conjunto de datos, las industrias como observaciones se representan como puntos rojos en el espacio bidimensional, mientras que las líneas azules representan la contribución de cada variable a los componentes principales en cada dimensión.

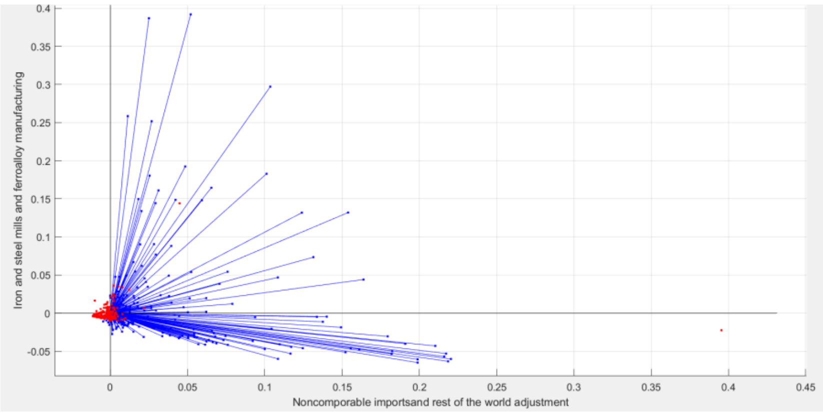


*Figura 2 (repetida): PCA de la muestra.*

A través del PCA realizado en toda la base de datos, se reveló que 'Noncomparable imports and rest of the world adjust' son la variable más dominante, explicando el 10.27% de la variabilidad total. Se descubrió que esta representación de la variabilidad no era una tarea fácil, lo cual se explica fuertemente por la gran dependencia de la economía en esta industria. Además de este resultado relevante, fue difícil discernir cualquier otra información en esta compleja economía, por lo que fue crucial llevar a cabo el PCA sectorial.

Esto consiste en realizar PCA en las aglomeraciones de industrias por sí mismas. Según el NAICS, las industrias pueden agregarse aún más en sectores, basándose en sus criterios. Se esperaría que las industrias en un mismo sector estén fuertemente interconectadas. El PCA se llevará a cabo en un sector que actuará como objetos y todas las industrias actuarán como variables.

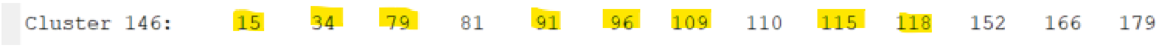
De esto se extrajo la siguiente información. La mayoría de los sectores sienten poco o ningún efecto de las industrias fuera de su sector, especialmente Finanzas y Seguros, y Atención Médica. El único sector que se vio fuertemente afectado por los cambios en la demanda fuera de su sector fue Industrias Especiales, que incluye los dos componentes principales del PCA general.

### K-means

Este método implica encontrar clusters no superpuestos y sus centros en un conjunto de datos no etiquetados como el que se tiene. Las decisiones para tomar para el diseño son el número de clusters que se desea obtener de los datos y qué método usar como criterio de agrupamiento.

Para encontrar este valor se realizó un análisis del Método del Codo. Se basa en la minimización de la Suma de Cuadrados Dentro del Cluster (WCSS), que es la suma de las distancias al cuadrado entre cada punto y el centroide de su cluster. A través de dos métodos para calcular el k óptimo, los datos se representaron óptimamente mediante 180 clusters.

Al agrupar basándose en este valor de k y a través de la distancia euclidiana, se llegó a una inconclusión ya que la composición del cluster, con más de un elemento, varió a través de las iteraciones. Solo algunos valores parecían repetirse a pesar de ellas.



```
Cluster 146:     15    34    79    81    91    96   109   110   115   118   152   166   179
```

*Figura 12: Composición del cluster compuesto, con repeticiones destacadas*.

Las industrias marcadas en amarillo fueron las que aparecieron conssistentemente. Algunos argumentos pueden explicar las relaciones desde diferentes puntos de vista. Como la similitud sectorial entre Construcción (15) y Bienes Raíces (118), debido a la relación simbiótica que comparten. Las relaciones entre estas industrias parecen ser claras bilateralmente, pero es extremadamente difícil encontrar un nexo común entre todas estas industrias debido a la complejidad de la imagen. Estas son las únicas relaciones que destacan en una economía tan compleja donde todos los objetos parecen no compartir similitudes con otros.

Al realizar el K-means para el valor más bajo posible para ver cuál es la industria más diferente, Noncomparable imports and rest of the world adjustment fue la primera en formar su propio cluster. Esto se debe a su dependencia en los cambios de demanda de toda la economía.

## CONCLUSIONES Y FUTUROS TRABAJOS

El PCA destacó la existencia de algunas industrias clave que influyen en la producción del resto, como Noncomparable imports and rest of the world adjustment, y a Steel Mills and Ferroalloy Manufacturing en el sector de Industrias Especiales. El análisis de K-means destacó la complejidad de las relaciones de producción, con un gran número de clusters como composición óptima que indican la contribución única de cada industria a la variabilidad económica. Esta afirmación también se respalda por el hecho de que los miembros del cluster con más de una industria variaron con las iteraciones.

Debido a la complejidad de la economía revelada, las políticas deberían considerar las interdependencias entre industrias en un mismo sector. Un ejemplo sería incentivar la demanda de las Compañías de Seguros para afectar el sector de la Vivienda. Las relaciones simbióticas, como esta entre la Construcción y los Bienes Raíces, sugieren que las políticas que afectan a una podrían tener un efecto dominó en la otra.

Este análisis ha encontrado algunas limitaciones. Como los resultados del método del codo y la puntuación de la silueta, los cuales fueron inconsistentes debido a la complejidad de los datos. La reducción del PCA a dos dimensiones puede llevar a una pérdida de información, especialmente con una variabilidad tan dispersa. El enfoque del estudio en un solo año y la antigüedad de los datos limita su aplicabilidad a las condiciones económicas actuales, pero no poseerá informes reales sobre el comportamiento de las industrias en esos respectivos años.

Para futuras investigaciones, se debería considerar el análisis de economías más pequeñas y simples para identificar patrones que puedan aplicarse a la economía estadounidense. Incorporar variables como los efectos de la inflación en la producción de la industria podría proporcionar informes más completos. Probar modelos predictivos podría responder qué interrelaciones encontradas en este análisis pueden aplicarse a la formulación efectiva de políticas económicas.

# *Index*

***Chapter 1. Introduction*** **5**

1.1 Objectives 6

1.2 Input-Output matrices and Leontief Model 6

1.3 Machine Learning 8

1.4 Structure of the project 9

***Chapter 2. State of play*** **10**

***Chapter 3. Technology description*** **13**

3.1 Leontief Matrix 13

3.2 Matlab 14

*3.2.1 Principal component analysis 14*

*3.2.1 K-means clustering 15*

***Chapter 4: Development*** **16**

4.1 Methodology 16

4.2 Taken decisions 17

4.3 Principal component analysis: 18

*4.3.1 Agriculture, Forestry, Fishing, and Hunting 22*

*4.3.2 Mining 23*

*4.3.3 Finance and Insurance 24*

*4.3.4 Health care and Social Assistance 25*

*4.3.5 Special Industries 26*

4.4 K-Means Analysis 27

*4.4.1 Optimal Number of Clusters 28*

*4.4.2. Squared Euclidian distance: 31*

***Chapter 5: Results and Conclusion*** **35**

5.1 Principal Component Analysis (PCA) 35

5.3 Summary and findings 37

5.4 Implications on Economic Policy 37

5.5 Limitations and Challenges 38

5.6 Future Research Directions ........................................................................................ 39

***Chapter 6. References*** ***........................................................................................ 40***

***ANNEX I: LIST OF ANALYZED INDUSTRIES .......................................................... 43***

***ANNEX II: MATLAB CODE........................................................................................ 48***

# *Figure Index*

Figure 12: Composition of composite cluster, with highlighted repetitions ....................... 11

Figure 1: Example of Input-Output Table [3] ......................................................................... 7

Figure 2: PCA of sample ......................................................................................................... 19

Figure 3: zoom on the most contributing variables on x axes. ........................................... 20

Figure 4: zoom  on the most contributing variables on  y axes. ......................................... 21

Figure 5: pca on Agriculture, Forestry, Fishing, and Hunting industries. ........................... 22

Figure 6: pca on mining industries. ........................................................................................ 23

Figure 7: pca on Finance and Insurance .................................................................................. 24

Figure 8: pca on HEALTH CARE AND SOCIAL ASSISTANCE. ................................... 25

Figure 9: pca on SPECIAL INDUSTRIES............................................................................. 26

Figure 10: Elbow method on data with 10 repetitions. ....................................................... 29

Figure 11: Silhouette analysis for optimal number of clusters. ........................................... 30

*Figure 12: Composition of composite cluster, with highlighted repetitions*................ 31

Figure 13: K-means clustering on the economy for k=180 ................................................ 32

Figure 14: K-means clustering on the economy for k=180 ................................................ 33

# *Table Index*

*Table 1: industries agglomerations or sectors.*

# CHAPTER 1. INTRODUCTION

In modern economics understanding interactions between its parts is crucial for effective policy making and the further development of its situation. Input-output matrices, developed by Wassily Leontief have been a fundamental tool for analysing the effects that those interactions may have on shifts in output. The problem arises when economies begin to escalate becoming more difficult to study, it must be complemented by advanced analysing tools such as machine learning. This study focuses on integrating those techniques, specifically clustering, to find those key elements of understanding that may be hidden from the human eye.

The project will be carried out with the objective of understanding in depth how the different sectors of the economy interact with each other, beyond the human perceptions that may be held in advance. Using machine learning, learning from the results and failures that this process can obtain. Taking advantage of the opportunities offered by these tools to analyse the intersectoral relationships of an economy, without being limited by the complexity of the model or human preconceptions.

The aim is to discover groupings and associations between sectors, which may not be evident to the human eye, but which allow a better understanding of the economy and the government measures to be taken. To do this, we want to find the most effective economic measures to affect either the production of a group of sectors, employment, or investment.

The aim is to find the limitations in human observation and in the analysis of how the different elements of production are grouped, avoiding prejudices that may arise due to preconceptions of the behaviour of the industry sectors. It will also aim to learn about the potential, restrictions and limitations of the tools offered by machine learning. By putting into practice the use of principal components and clustering to reduce the problem to two dimensions, the main goal will be to study the validity of the representations and which alternative may be more beneficial to understand the associations between sectors.

## *1.1 OBJECTIVES*

1.1.1 Analyse the Input-Output matrix of the United States

Carrying out the Leontief model, which will allow obtaining the forward and backward relationships of the economic system, understanding the matrix relationships in the process.

1.1.2 Evaluate the effectiveness and applicability of machine learning to the topic.

Whether complex economies are suitable for this kind of tools. Considering both the effectiveness of the process and the traceability of the results obtained to the original situation.

1.1.3 Identify patterns and clusters of industries that reflect change in the economy.

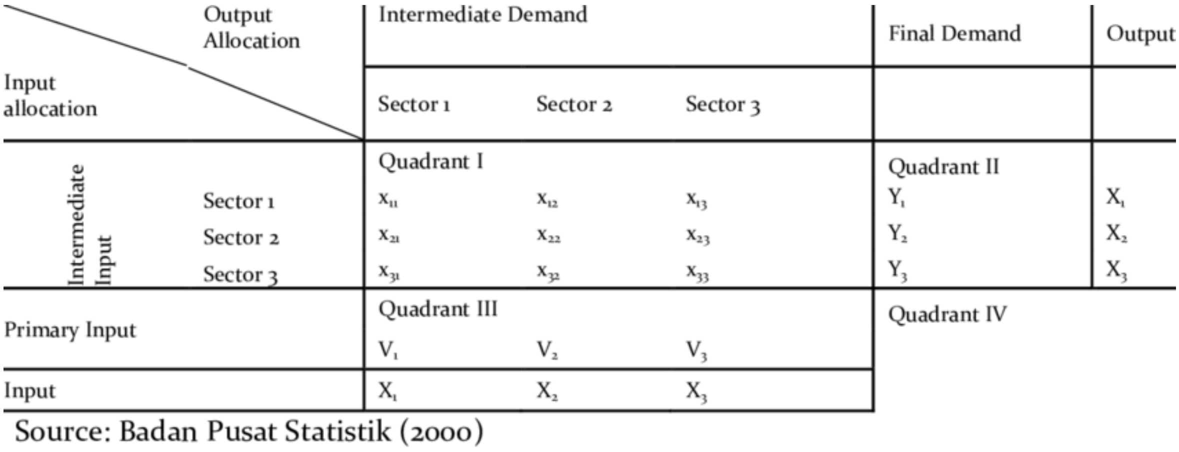
Whether there are indicators that can help predict behaviour in the sectors of the economy. Through the variation of demand and added values, which are government dependent, in order to create recommendations for economic policies.

1.1.4 Analyse the possibility of developing predictive models.

Based on the clustering result, contemplate the validity of developing a predictive model. This will involve developing a supervised model.

## *1.2 INPUT-OUTPUT MATRICES AND LEONTIEF MODEL*

Input-output matrices, also known as IOTs, were developed by Wassily Leontief in 1936. He builds from the “Tableau Economique” of Francois Quesnay [1] and “Elements d’ Economie Politique Pure” de Leon Walras [2], for the economic theory of general equilibrium. He received a Nobel Prize in Economic Science in 1973 due to this achievement. His theories have contributed to the understanding of economic development and movement, they are used openly by governments to understand the areas of promotion and causality in the different industry sectors. His model analyzes the interdependence between the different industries of an economy focusing mainly on the relationship between the resources generated by an industry and how these are consumed by others. More specifically, the model, represented by a matrix, contains in its rows the production of each industry and in its columns the consumption of goods from industries to carry out their production. The most common table to represent this is the "USE" table, which is very reminiscent of a recipe since it indicates the elements required by an industry to carry out its production, an example of this table would be Figure 1 shown below:



| Input allocation | Output Allocation | Intermediate Demand | | | Final Demand | Output |
|---|---|---|---|---|---|---|
| | | Sector 1 | Sector 2 | Sector 3 | | |
| Intermediate Input | | Quadrant I | | | Quadrant II | |
| | Sector 1 | $x_{11}$ | $x_{12}$ | $x_{13}$ | $Y_1$ | $X_1$ |
| | Sector 2 | $x_{21}$ | $x_{22}$ | $x_{23}$ | $Y_2$ | $X_2$ |
| | Sector 3 | $x_{31}$ | $x_{32}$ | $x_{33}$ | $Y_3$ | $X_3$ |
| Primary Input | | Quadrant III | | | Quadrant IV | |
| | | $V_1$ | $V_2$ | $V_3$ | | |
| Input | | $X_1$ | $X_2$ | $X_3$ | | |

Source: Badan Pusat Statistik (2000)

*Figure 2: Example of Input-Output Table [3]*

The figure shows quadrant one as the intermediate relations, which are the relations between industries. Quadrant two are the forms of demand, both consumption and investment, as well as government. Quadrant three are the forms of added value in the production, an example would be the use of capital or the payment of taxes in its production. By last, the last row and column represent, respectively, the components and production finals. These should coincide, conforming to Leon Walras's imposed equilibrium premise [4].

Leontief's model is of volumes not of prices, it considers that the price varies in perfect elasticity with production [5]. It will allow us to obtain a relationship between the changes in demand of a sector and changes in production of the economy. This model can be seen as a representative of production multipliers, explaining how changes in demand affect final production [6]. The analysis process of these relationships is very complex due to the extension of sectors, which is why it is usually resorted to a simple ordering between sectors and then analysing the elements separately. This process is not beneficial since it can ignore relationships between sectors by reducing them to a simple numerical ordering. Today there is a method to make these groupings that includes relationships between sectors with much better precision, this is machine learning.

## 1.3 *MACHINE LEARNING*

Machine learning or statistical learning may sound like a new concept, but the field has been developing since the beginning of the 19th century. The method development of the least squares by Legendre and Gauss, which could be considered the earliest form of linear regression, could be used for quantitative and qualitative predictions.

In the 1980s, thanks to the development of computational technologies non-linear methods became viable. Regression trees were introduced, and since then many other methods. A powerful tool was rediscovered due to the possibilities it seemed to offer. Many other disciplines emerged, such as clustering. [7]

Clustering is a methodology, machine learning, where the algorithm analyses data to find patterns, without the need for human intervention. This methodology earns its name from the fact that the results cannot be contrasted. It is not a formula that when used against a new database will give, the expected results. The objective of using this tool is to find associations or groupings between elements without the need for human intervention. Whether to prevent it from contaminating the associations with human preconceptions, that the volume of data is greater than what is humanly analysable or because no apparent relationship is found between concepts.

## *1.4 STRUCTURE OF THE PROJECT*

The thesis begins with a historical background of the Leontief model and its impacts on modern economic analysis, providing examples of its applications. It then describes the technology used in the project, including the Leontief Matrix and the implementation of MATLAB code. The methodology applied for the analysis is detailed, including the decisions made during development and their justifications. The applications of PCA and K-means clustering are thoroughly examined.

The results and conclusions obtained from the study are presented, summarizing the findings, including both PCA and K-means results. The implications for policymaking are discussed, highlighting how the results may be applicable. The thesis also addresses the limitations and challenges faced during the study and suggests future work opportunities based on the findings.

# CHAPTER 2. STATE OF PLAY

Wassily Leontief back in the mid-20th century, developed the first empirical implementation of a general equilibrium model for the economy. This allowed deeper analysis on the economy thank to the possibilities it offered. It surged as a practical, problem-solving discipline based on the collection and processing of data that led to a simplified analytical tool. Its impacts on modern economy have been countless, introducing a new way to look at the economy and its behaviour [8]. This is the reason why it is used and mixed in several studies such as combining it with the Armington assumption, to focus on the domestic and imported output. By distinctively separating domestic production from the rest, it is claimed that Leontief multipliers are more precisely calculated [9]. This adaptation will not be carried out due to the added layer of complexity and for a more globally accepted approach.

IOTs have been around for almost a century, allowing the in-depth study on the matter and its potential uses and which has allowed many governments, individuals and organisations to analyse the economy through their use on numerous occasions. An example would be the Andalusian government, which carried out an analysis of the structural features of its economy [15]or the analysis of the Greek economy after the crisis [16]. In this last one, the multipliers of production, investment and employment, and their respective elasticities, are analysed. After calculating the mentioned parameters, a ranking is carried out by parameter and the sectors are compared based on the position in the ranking of the parameters. This method allows us to analyse certain sectors within the economy and group together some sectors that share clear hierarchical relationships between categories. This forces them to analyse only certain areas of the economy since the structure is excessively complex, such as elements that appear systematically at the top or bottom of their classifiers.

Some studies in the matter are not carried once to gasp a glimpse at the state of the interrelations but periodically so that tendencies and developments can be seen. This is the case of the Office of Financial Management of the Washington State that carried out this type of study every five years till 2012. They developed a model with 52 sectors to later calculate the influence of employment and their income had on the economy. This was done to figure out how the movement of output could be influenced by those two categories [12]. As it can be seen the model has a huge power of adaptability, multipliers may vary due to the scope the study may want to apply.

It is a common trace in this type of analysis to have to reduce the object in order to obtain clear conclusions from the analysis, such as focusing on a few indicators and ignoring the rest. This leads to the waste of information because of its complexity. Machine learning provides the opportunity to leverage information in greater depth. An example of using information through these tools would be the academic publication the analysis on academic performance of students [17]. In this work, clustering is used to group students based on their grades, showing them, using two-dimensional representations. Next, it relates behaviours based on their grades, and using these relationships obtained, it tries to predict the students' results at the end of the semester. The tool used fits perfectly with the problem to be solved, since we want to avoid human intervention, while trying to understand relationships that may not be obvious at first glance.

Many intellectuals rapidly understood this advantage and applied it to their respective sectors of the economy. Applied economics to the technology started back in 1974, but it is not until 1984 that questions were asked on the possibilities of the tool by Wang in his respective study on the scoring models obtained [14]. More recently Aaron Smalter Hall applied deep learning techniques to find the optimal forecasting model for the unemployment rate. Through testing the errors the predictive models obtained through machine learning and adjusting the parameters he obtains an optimal solution [15].

Another example of students taking the lead on the applications of machine learning was by Cornell University students. They study examines how the changes in the interest rates made by the FED affect the returns if fixed income and equity funds. It combines gradient boosting and linear regression, to provide indicators on how the funds returns react to those changes, so that opportunities can be created both in future studies and political decisions [16]. With a similar intention the International Monetary Fund carried out the following study, by utilizing dynamic factor models and machine learning algorithms to predict the GDP growth. It proved to be an effective method to integrating large volumes of data, such as Google searches and air quality indicators, so that a policy response can be obtained in times of uncertainty [17].

Because of the longevity of both resources used in this paper, they have been used and studied on many occasions, proving to be a useful and powerful tool for analysing, policy making and developing predictive models. Although in this study only the first two will be covered, it opens the door for future studies on the predictive validity models generated with the information obtained. This model will focus on the structure of the economy based on their main drivers of output from within the industries and how they aggregate based on them, thanks to the Leontief multipliers. This will be carried out so that a clearer picture of the changes in output from a demand perspective can be obtained for policy making, studying and predictive models to depart from.

# CHAPTER 3. TECHNOLOGY DESCRIPTION

## 3.1 LEONTIEF MATRIX

Leontief model, also known as "*interdependence coefficients matrix*", is a tool that represents the outputs multipliers [8]. These multipliers are calculated through algebraic operations, in which i and j represent rows and columns respectively. Initially the intermediate transactions between sectors are need (Z), represented by quadrant 1 of Figure 2.These need to be divided by the total industry output (X), which is calculating by adding all elements of each row, to obtain the direct element matrix:

$$A = [a_{ij}] = [\frac{Z_{ij}}{X_i}]$$

*Equation 3: Direct element matrix*

Which represent the coefficients of the intermediate transactions of the industries, based on their total output. Therefore, the economic model could be expressed with the following equation:

$$X = AX + D$$

*Equation 4: Open Economy Representation*

Where X represents the total output and D represents the total demand for that industry. This equation can be simplified furthermore, by factoring out the total output, obtaining Leontief's inverse (L):

$$L = (I - A)^{-1}$$

*Equation 5: Leontief Matrix*

$$X = LD$$

*Equation 6: Leontief's Open Model*

In the Leontief Matrix, also called "Leontief's Inverse", the elements of their rows represent the effects changes in demand of the economy may have on the output of an industry. If these are accumulated, the total potential effects on a sector's output are isolated from the rest of the sectors of the economy. This matrix is base for the model and is sometimes seen as the seller's point of view, due to how the potential impacts on the sectors are represented [5]. This is the primary tool and point of view, to be used in order to classify the sectors of the economy.

## 3.2 MATLAB

### 3.2.1 PRINCIPAL COMPONENT ANALYSIS

It is a powerful tool that permits the summarization of correlation between variables of a data set. This is done by representing and explaining the maximum variability of the set, with the minimum number of variables.

The following MATLAB function is used:

*[coeff,score,~,~,explained,~] = pca(A)*

*Equation 7: PCA function Matlab*

*Coeff:* each column represents the coefficients for each principal component. They are ordered by their capability to explain the total variability of the data set.

*Score:* are the representation of the variables in the principal component space.

*Explained:* returns the percentage of the total variance of the model explained by each principal component. [9]

### 3.2.1 K-MEANS CLUSTERING

It is a method that permits portioning the data set into K, non-overlapping, groups also called clusters. Each point of data is belonging to a group with the nearest mean, while generating the smallest variance in the group. [7]

$$[idx, C, sumd, D] \ = \ kmeans(A, k, 'Replicates', 10)$$

*Equation 8: kmeans function Matlab.*

Inputs:

*k*: number of clusters desired.

*'replicates'*: in order to avoid local minimums, it carries out the algorithm 10 times.

Outputs:

*idx*: Cluster indices assigned to each observation.

*C*: Coordinates of the cluster centroids.

*sumd*: Sum of squared distances from each point to its centroid.

*D*: Distances from each point to each centroid. [10]

# CHAPTER 4: DEVELOPMENT

## 4.1 METHODOLOGY

Initially, the Leontief model must be obtained from the IOTs. Having the USE table, quadrant one must be transformed into the direct requirements matrix A. This is achieved by dividing all intersectoral inputs between production by sector. We will use this matrix to calculate the multiples of production, investment and employment; through matrix operations, represented by Equation 3, Equation 4, Equation 5 and Equation 6.

Once the matrix is created, PCA or "principal components analysis" will be used; it will be validated by the percentage of the variability of the economy represented by the PCA. An image of the most representative two-dimensional space of the variability will be obtained. Although it will be a valid graph, the representation and information highlighted will be put to the test. It may be the best representation obtained through that tool but it does not necessarily mean that the information is clear or easy to discern. The industries and their scores will be represented to analyse which changes of demands in industries have similar directions of the variability represented. This is another method of obtaining info from the two-dimensional space since some industries' changes of output will lightly be represented in two dimensions when the data has 192 dimensions.

The use of clusters will also be considered, by limiting the number of groups. The objective will be to find out whether some industries behave similarly to similar changes in demand. To find out if industries can be aggregated to be considered or to simplify the developments on monetary policies. The first step will be to find out the optimal number of clusters so that the optimal complexity is found. This point will find the optimal amount of variance represented while avoiding over-saturation of clusters and becoming a biased representation. It is the fine line between having a too complex model with a small number of clusters where interrelations are hard to see and an oversimplified clustering that reports little to no information.

Several methods will be carried out to get an idea of the optimal number of clusters the model requires. To later see on what that value means and how industries behave under that level of clustering. Clustering techniques will be applied based on what to consider as the optimal distance relation between variables and how representative the technique seems to be.

Finally, a reflection on the information gathered, the possibilities it offers and its limitations will be carried out. It is expected to have a better understanding of the economy and the functioning of the used tools, as of answering some of the questions and objectives presented in the introduction. Future works on the matter and different approaches to be taken will also be highlighted when all information is gathered and analysed.

## *4.2 TAKEN DECISIONS*

For the building of the Leontief matrix, the data was obtained from the Bureau of Labor and Statistics (BLS), where a "USE" matrix of real values was obtained. Therefore, the analysis will be carried out of the United States economy, this is due to the amount of data available and the complexity it may offer. BLS is a recognised and reliable source of information that is constantly monitored. This database also offers a US economy that is largely fragmented which allows the highlighting of inter-area relations, that would go under the radar otherwise.

The decision to use real values was based on the avoidance of the effects that inflation and other price-distorting factors may have on the analysis. To further focus on the effects changes in demand may have on individual productions and the economy. The use of a "USE" matrix is more of a requirement than a decision. It puts more emphasis on consumption rather than production, which means that the added values of the products are not included in each variable and its separated as a different object which can be avoided in the study. This can allow the inter-industrial relations to be studied as cleanly as possible.

The data is divided into 192 individual industries that are grouped by BLS into the following:

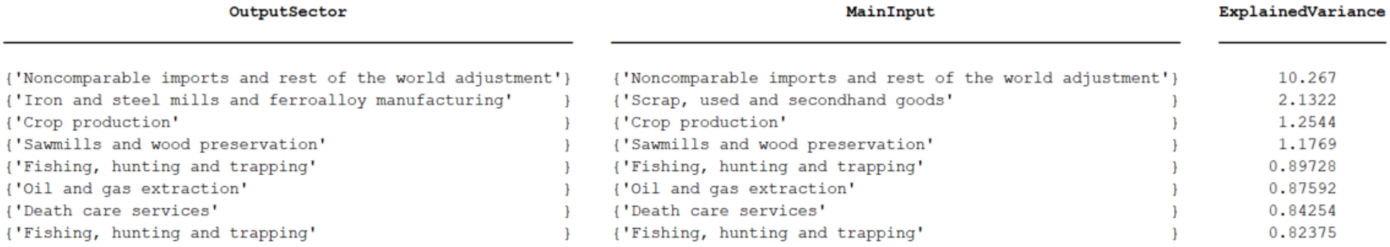
*Table 2: industries agglomerations or sectors.*

'Agriculture, forestry, fishing and hunting'
'Mining'
'Utilities '
'Construction'
'Manufacturing'
'Wholesale trade'
'Retail trade'
'Transportation and warehousing'
'Information'
'Finance and insurance'
'Real estate and rental and leasing'
'Professional, scientific, and technical services'
'Management of companies and enterprises'
Administrative and support and waste management'
'Educational services'
'Health care and social assistance'
'Arts, entertainment, and recreation'
'Accommodation and food services'
'Other services (except public administration)'
'Government'
'Special industries'

This is done according to the North American Industrial Classification System (NAICS) 2022. Which groups based on production processes, input requirements output characteristics, market orientation and technological similarities.

## 4.3 PRINCIPAL COMPONENT ANALYSIS:

An initial PCA will be carried out in the whole Leontief sample, using the function explained in Equation 7. According to the results obtained this are the variables that explain most of the variance of the sample:



```
                   OutputSector                                        MainInput                        ExplainedVariance
    ________________________________________________    ________________________________________________    _________________

    {'Noncomparable imports and rest of the world adjustment'}    {'Noncomparable imports and rest of the world adjustment'}         10.267
    {'Iron and steel mills and ferroalloy manufacturing'      }    {'Scrap, used and secondhand goods'                     }         2.1322
    {'Crop production'                                        }    {'Crop production'                                      }         1.2544
    {'Sawmills and wood preservation'                         }    {'Sawmills and wood preservation'                       }         1.1769
    {'Fishing, hunting and trapping'                          }    {'Fishing, hunting and trapping'                        }        0.89728
    {'Oil and gas extraction'                                 }    {'Oil and gas extraction'                               }        0.87592
    {'Death care services'                                    }    {'Death care services'                                  }        0.84254
    {'Fishing, hunting and trapping'                          }    {'Fishing, hunting and trapping'                        }        0.82375
```

Table 3: main variance representatives and their main inputs.

As can be seen, the only dominant variable seems to be 'Noncomparable imports and rest of the world adjustment' which accounts for 10.27% of the total variance of the sample. This will indicate that the industry as an input has the major influence on the economy's production as a whole, this is probably due to the heavy dependence the United States has on imported resources, such as semiconductors. We can also see that the industry mostly relies on its own services and products in its production process as the main input is itself.

After that variables seem to represent little to nothing of the total variance of the sample and their main inputs seem to be themselves. This means their production is mostly dependent on themselves, which could be because of them being vertically integrated and their search to avoid dependence on others. An exception to this could be the second variable 'Iron and steel mills and ferroalloy manufacturing' which is mostly dependent on the 'Scrap, used and second-hand goods', representing 2.13% of the total variance. This could be because the availability of scraps makes iron production so much cheaper, that it has a bigger impact on the sector than its own production which is probably much more expensive due to the manufacturing processes it requires. This could also be due to the increasing availability of scrap materials especially since we live in a recycling culture.

The data can later be represented in the two-dimensional space, thew observations represented by red dots, and the other relations or variables represented by blue lines:

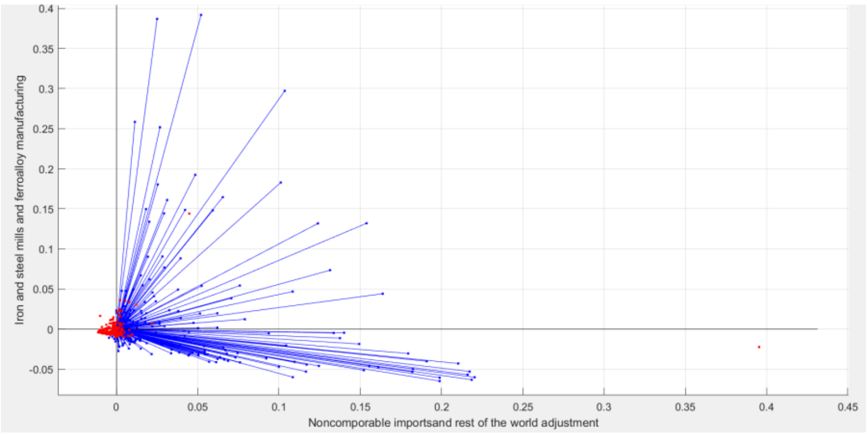


*Figure 3: PCA of sample*

Observations are represented in the space by the red dots using the scores obtained for the two most representative inputs which were previously mentioned in the Table 3. As we can see in the build space, observations have similar values, this is why we see most of the observations aggregated around the origin. This may suggest that the variance is so dispersed between the variables that the representation on this surface is only representative of the two observations that are also the main inputs. These are the red dot on (0.4, -0.02) coordinates representing Noncomparable Imports and Rest of the World Adjustment and on (0.04, 0.14) representing Scraps, Used and Second hand Goods.

The blue lines represent the coefficients of the original variables on the two-dimensional space of the first two inputs. It tells us how much of the original variable contributes to each principal component. Longer lines indicate that it contributes more, and their specific directions tell us what principal component contributes more. Taking a closer look on the 1st principal component most contributing variables we can see:

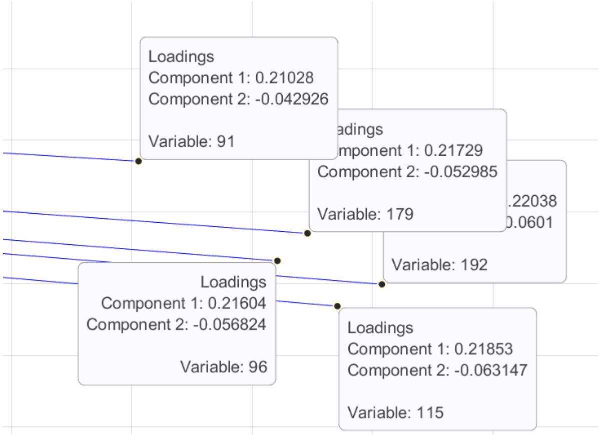


*Figure 4: zoom on the most contributing variables on x axes.*

Variable 192, Noncomparable Imports and Rest of the World Adjustment, seems to be the furthest right, which is no surprise after the previously stated information, but it seems to be followed by many other sectors. Securities and investment (115), and Federal Defense Spending (179), seem to be the closest. The fist could potentially be explained by the fact that economic stability and growth is crucial for the capital allocation of the country. The spending in defense seems to also be extremely correlated with both variables, due to the large movement of money United States has in this category it may incentive the external investment of other countries in search of being part of this investment. Wholesale trade (91) and Air Transportation (96) are easier to explain since many of the imports are typically linked with supply chain requirements for the movement of goods. This normally includes one or both of the previously mentioned industries.

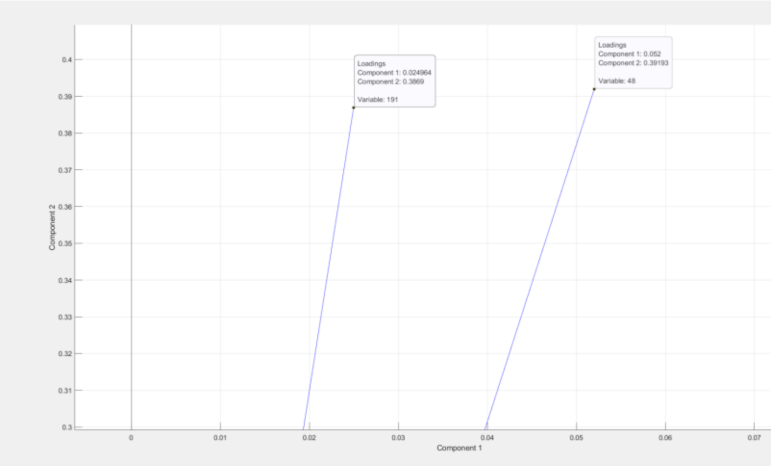


*Figure 5: zoom on the most contributing variables on y axes.*

The most contributing variables on the y axes seem to be less complex, having only two variables. They have the highest loadings on the axis meaning that they are the major contributors to the second component of variance. Scraps, Used and Second hand Goods (191) is easily explained as it is the second component that explains the variance of the sample. Iron and Steel Mills and Ferroalloy Manufacturing (48) is the industry responsible for producing alloys involving iron, it is understandable that is highly correlated with the availability of scraps as input of their production. In an industrially developed economy, a heavy dependence on these two sectors was expected, as it its crucial for the functioning of many other industries.

The problem that arises after this analysis is that there seems to be too much information on a single graph. Due to the complexity of the model representations can become saturated with information, not allowing the less impactful relations to appear. A feasible solution could be to analyse the agglomerations of industries on their own. Maintaining all industries as variables and sources of variance, but analysing only the members of the agglomeration, permitting a more comprehensive analysis. Key drivers of variance may appear in the close look approach. Although members of the same agglomeration would be expected to have the highest influence on the pca, this could not always be the case in some industries. It could also happen that an industry that is not typically considered as part of the agglomeration has a higher influence than what had previously been expected. Therefore, a PCA analysis will be carried out in some of the most important agglomerations, showing a new graph where the y axes represent the total variance represented, in the specified agglomeration selected in the title, for each individual industry represented in the x axes as a number.

### 4.3.1 AGRICULTURE, FORESTRY, FISHING, AND HUNTING

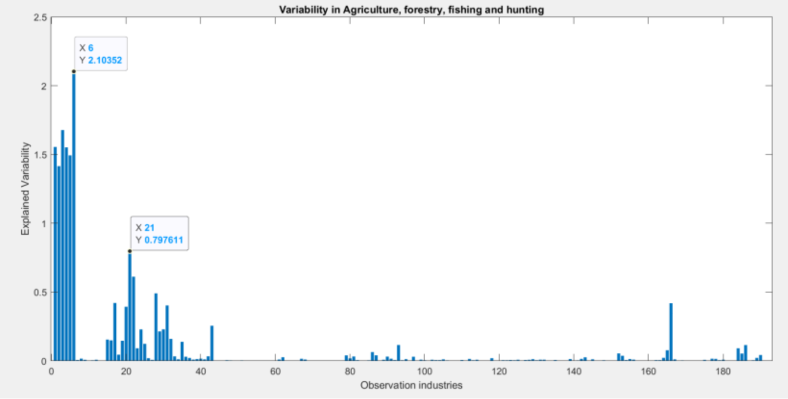


*Figure 6: pca on Agriculture, Forestry, Fishing, and Hunting industries.*

The first agglomeration, Agriculture, Forestry, Fishing, and Hunting, is represented, including industries 1 to 6.

We can see that the interrelation in the agglomeration is very high, as the variance of the demand of the first 6 industries has a high effect on the production of all industries. There seems to be a significant peak for Support Activities for Agriculture and Forestry (6), this could be because as a supportive activity has a high impact on all the other productions, whereas the others do not affect others as much. So there is high dependence on the other industries' production for those supportive activities to make their production easier and cheaper, it is an industry that affects the other's production significantly so it can be a good way to incentivise the production of the agglomeration.

There is also another peak that should be mentioned since it belongs to another agglomeration. Animal Slaughtering and Processing (21) demand seems to have a high impact on the variance of the agglomeration. This is no surprise because it could be considered the closest step in the supply chain between the agglomeration and the food industry. The changes in demand for animal processing will evidently require the increase of production of those industries in charge of hunting and fishing, with a special emphasis on fishing since only 7% of national consumption comes from aquaculture [18].

### 4.3.2 MINING

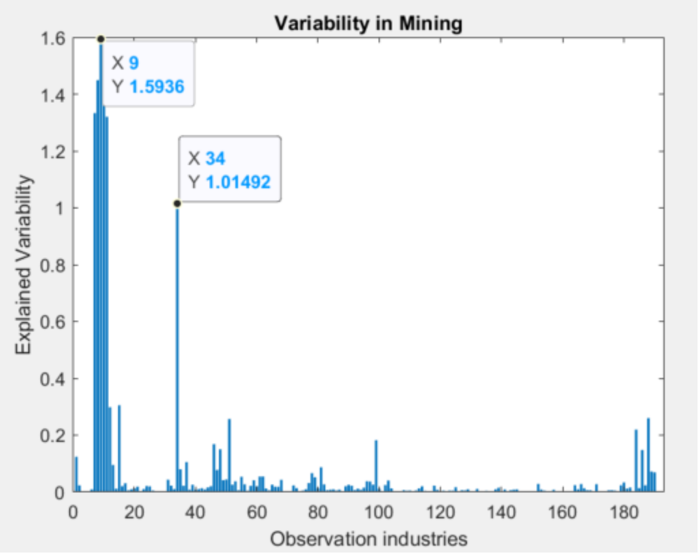


*Figure 7: pca on mining industries.*

Although we have seen that imports are the most crucial aspect of the US production economy, there seems to be a high dependence on steel scraps in order to develop the metal industry. It will be greatly impactful to understand how deeply those sectors affect the extraction of those metals on their land, this is why a PCA on the mining industries has been carried out to obtain the previous figure.

As is shown in the graph and could previously be expected the interdependence of the industries of the agglomeration is high. Oil and gas extraction (7), Coal mining (8), Metal ore mining (9), Nonmetallic mineral mining and quarrying (10) and Support activities for mining (11) seem to be strongly connected. Metal ore mining seems to be the element that explains the variability of the agglomeration the most, this is surprising because Support activities for mining is the industry that affects the rest the most. This is explained by the fact that the Metal ore mining's production is highly dependent on its own demand multiplier while the supply activities production is mostly dependent on the rest.

It is also obliged to mention the existence of another peak in the graph that seems to be outside of the agglomeration. This is Petroleum and coal products manufacturing (34), which is understandable. Changes in demand for petroleum demand will have a high impact on the mining production. It is feasibly that the extraction of petroleum requires the same capital infrastructure as the mining agglomeration, making it cheaper and allowing the further production at a similar cost.

### 4.3.3 FINANCE AND INSURANCE

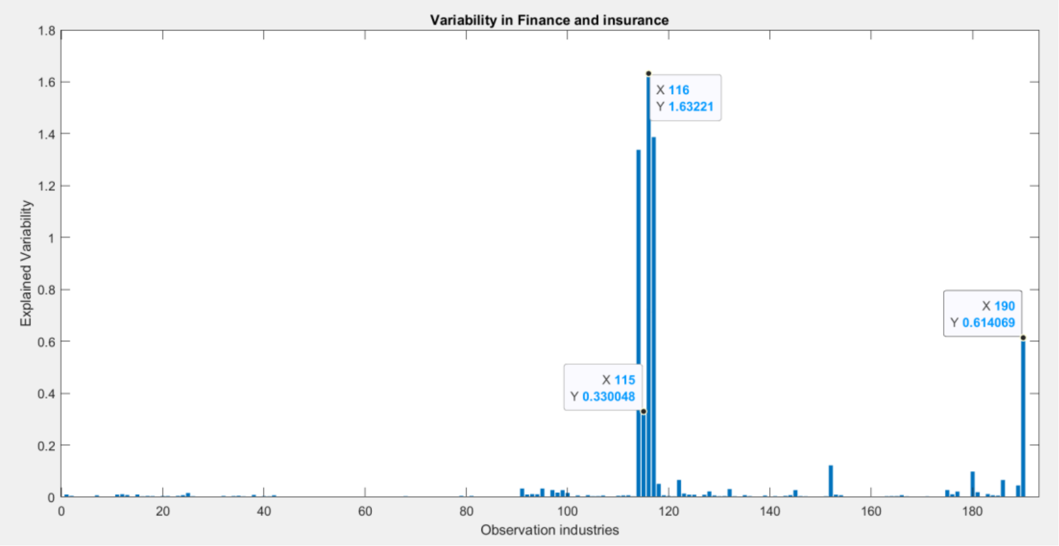


*Figure 8: pca on Finance and Insurance*

The highest variability represented by an industry seems to be Insurance Carriers and Related Activities (116), but it is closely followed by two others. Nevertheless, it seems to be a good option to affect the agglomerations production. This could be because it represents a crucial factor for managing financial activities, both to make and take them.

It is curious to see that there is a higher peak from the variability represented by an industry outside the agglomeration than by one of the groups. Owner-occupied Dwellings (190) have a higher significance in the finance and insurance sector than Securities, Commodity Contracts, Investments, and Related Activities (115). It should be noted that possibly due to the difference in the capital moved in both businesses, securities become less influential whereas dwelling that may require the movement of large amounts of capital from the financial sector highly incentivises their production. It is the only industry outside the sector that affects it, the reliance should be very high, and the activity must be remarkable.

### 4.3.4 HEALTH CARE AND SOCIAL ASSISTANCE

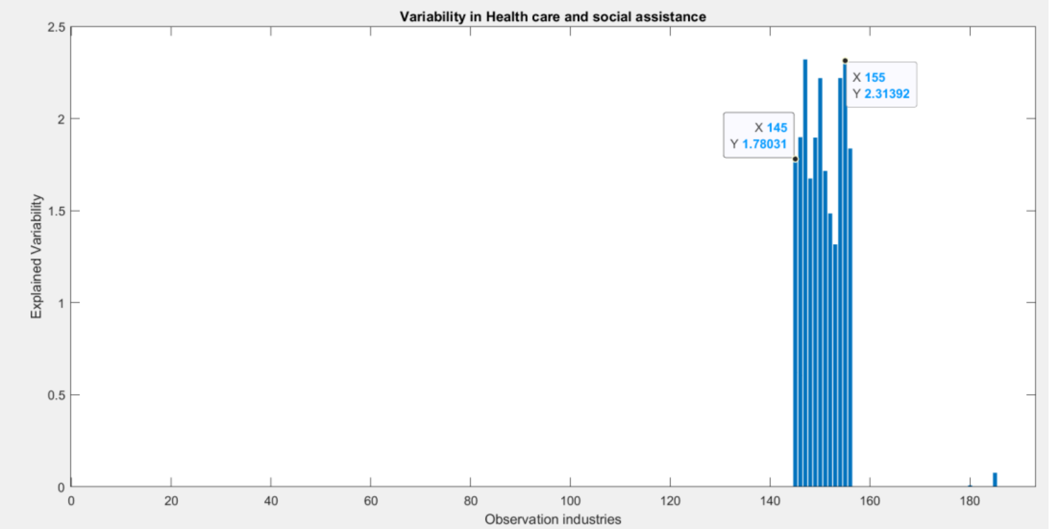


*Figure 9: pca on HEALTH CARE AND SOCIAL ASSISTANCE.*

This is a special sector due to the little to no variance industries have outside the sector. It is a self-sufficient and self-dependant sector. Having a peak for Community food and housing, emergency and other relief services, and vocational rehabilitation services (155) but being closely followed by others.

### 4.3.5 SPECIAL INDUSTRIES

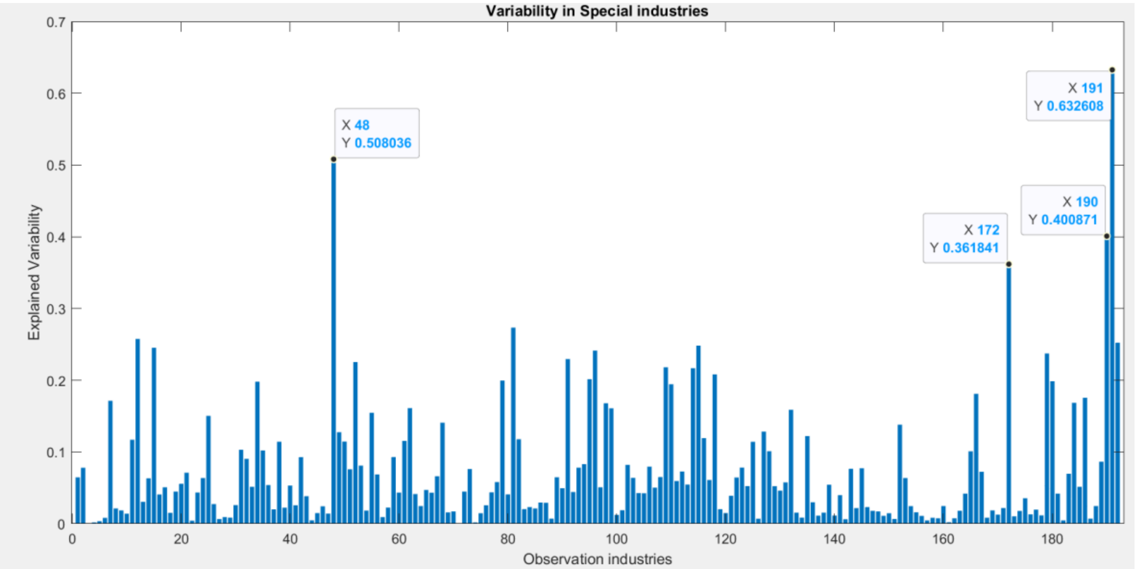


*Figure 10: pca on SPECIAL INDUSTRIES*

This graph shows the most movement of industries since more industries represent the variability of the sector. So not only does this sector contain the two industries that better explain the variability of the economy but also its production depends on many other demands. It is a key sector to both the economy and to surveillance on the state of it.

Death care services (172) although less represented than the other highlighted industries is a shocking industry to represent variability in a sector. In fact, there seems to be no apparent explanation or study on the internet. This is a perfect example of what was previously stated as one of the objectives of the project. There seems to be a strong correlation between the changes in demand for death care services and the changes in output for Owner-occupied dwellings (190), Scrap, used and second-hand goods (191) and Noncomparable imports and rest of the world adjustment (192). Perhaps it is a one time correlation or it may require further investigation.

Scrap, used and second-hand goods (191) have the highest variability explained in the sector. Closely followed by Iron and Steel Mills and Ferroalloy Manufacturing (48) which was the industry that better represented the change in production of industry 191 in the total PCA, so it's no surprise. As we mentioned previously these two industries seem to be strongly correlated, due to the demand for scraps and steel mills affecting strongly the production of their counterpart. What would not be expecting the low variability represented by the other two industries in the sectors Owner-occupied dwellings (190) and especially Noncomparable imports and rest of the world adjustment (192) which was the main component of variability in the total PCA. This means that the industry has a bigger impact on the changes of output of the global economy compared to the rest of the industries, rather than particularly in their own sector.

## *4.4 K-MEANS ANALYSIS*

As previously explained, this method involves the finding of non-overlapping clusters and their centres in a set of unlabelled data such as the one we have. The decisions to make for the design the number of clusters one desires to obtain from the data and what method to use as clustering criteria.

The objective of these two decisions is to find out the most representing number of clusters and the composition of those same clusters. This can be analysed through several methods that will later be analysed, but the objective is to minimise the differences between clusters without arriving at the redundant solution that each variable represents a cluster. To summarize the objective of this method is to obtain a result that represents the major differences in the data without going over the rest of the information.

This type of clustering works in the following way. Once the number of desired clusters has been specified, the centroids are settled in the data at random. All points are assigned to each centroid according to the distance to the centroids. Then the centroids are repositioned according to the selected criteria of distances. For example, the minimal square distance between points and the centroids or the minimal distance between points at the same cluster. Once centroids have changed positions, data points are reorganised again towards the closest cluster. This process is repeated until centroids remain at the same position despite iterations.

As it may appear this process stands out due to its computable simplicity which allows it to manage large groups of data.

The only problem is that the initial positioning of the centroids may lead to different final answers. This is because each start of the process may lead eventually to a local minimum, a point where centroids no longer need to vary their positions, but that does not specifically require them to form the clusters with less variance between points. The only solution to this problem is to carry out the process several times expecting to obtain the global optimum in one of the iterations..

### 4.4.1 Optimal Number of Clusters

There are several methods that allow the designer to decide on the optimal number of clusters to divide the data into, that are already build into MATLAB. These will be chosen based on the global acceptance as to being representative forms of evaluating the validity of a k-means distribution. This will be the elbow method and silhouette analysis.

- Elbow method:

It is a way to visualise the ideal number of clusters to carry out a k-means clustering. It is based on the minimisation of Within-Cluster Sum of Squares (WCSS), which is the sum of the squared distances between each point the centroid of its cluster [19]. Given the observations ($x_1$,$x_2$,..), k is the number of centroids and sets ($S_1$, $S_2$,…) as the clusters formed. Being $\mu_i$ the coordinates for the centroid the WCSS can be represented in the following equation:

$$arg\ min \sum_{i=1}^{k} \sum_{x \in S_i} \|x - \mu_i\|^2$$

*Equation 9: WCSS calculation [20].*

As it may have become evident, this measurement has a Achilles' heel. If the objective is to solely reduce its value to 0, then the optimal value would be for each point to form its own cluster. Therefore, the objective is to find out the value of k that makes the value of WCSS to verily change, in comparation to previous changes, when adding an extra cluster k+1. This value is better identified visually, by posting the minimal value of WCSS for each k number of clusters:

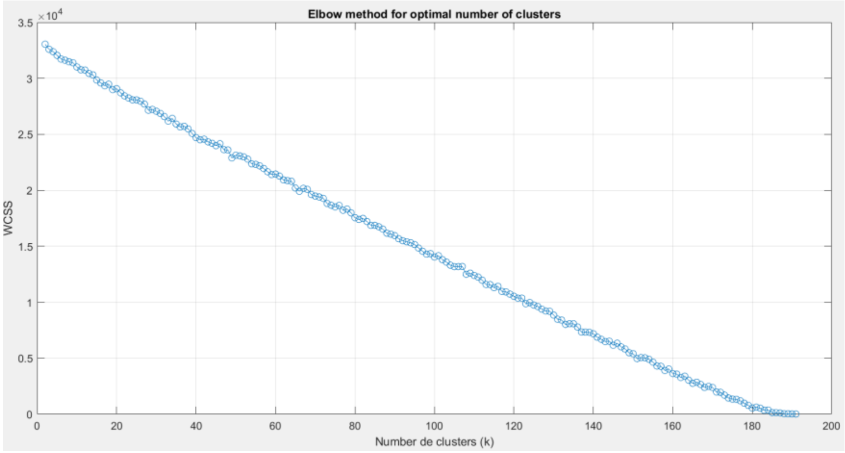


*Figure 11: Elbow method on data with 10 repetitions.*

This is not the way a good elbow method graph should look like, in fact it looks like an ankle. As the number of clusters increases the value of WCSS decreases lineally. In any case if a suitable answer could be obtained from this graph is that 180 seems to be the number of clusters, from which point WCSS does not decrease with the same steepness. Although it may not be a satisfactory answer, 180 centroids is the optimal number of clusters according to this method. It can also be valid that it is not a valid result and there is no optimal number of clusters.

The constant slope on the graph may indicate that to capture the behaviour of the economy towards changes in demand many clusters would be required. Meaning the industries' behaviour is diverse and complex. Luckily there are other methods that may allow us to find a lower value of optimal number of clusters.

- Silhouette analysis:

This method not only considers the distance between objects of a same cluster but also accounts for the distance between clusters, or how much distance there are between them, this is done to consider how evident is the differentiation.

The process is carried out in the following way. Firstly, for each object the average distance to each point of the same cluster is calculated ($a_i$). Secondly, for each object the minimal average distance to each other cluster is calculated ($b_i$). Thirdly, the silhouette score for each is obtained ($s_i$):

$$s_i = \frac{b_i - a_i}{max\{a_i, b_i\}}$$

*Equation 10: silhouette score for each object.*

This value can vary from 1 when the average distance between the points next to it is significantly smaller than the average distance to any other object in another cluster leading the silhouette score to be a division of $b_i$s. The other option occurs when the average distance between of a point to the members of a same cluster is significantly bigger than the average distance from that same point to the objects of other clusters. This will happen if clustering is conditioned or if not done correctly, but a score of 0 is viable and it may occur when distance between clusters is about the same as the one between members of the same cluster. The objective with this analysis is to obtain the number of clusters that offers the highest score, which will mean that the distance between members of same cluster s maximised and the distance with other clusters is minimised:

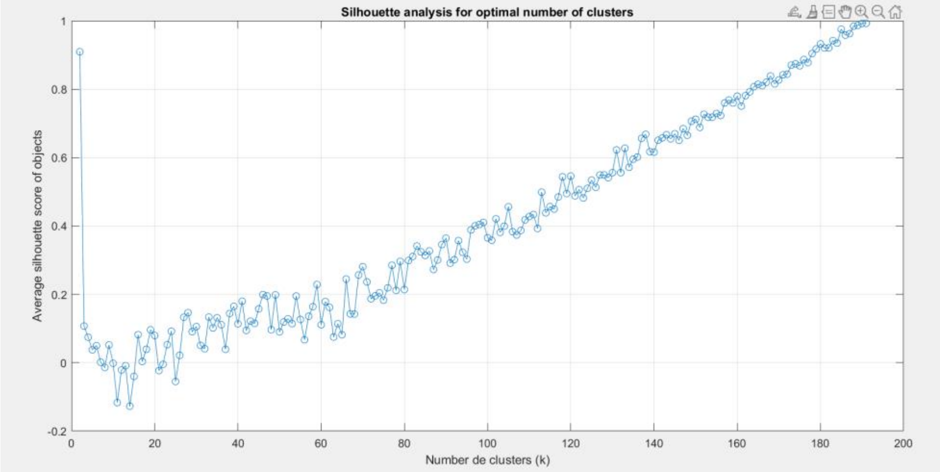


*Figure 12: Silhouette analysis for optimal number of clusters.*

This graph leads us to a similar conclusion than the elbow method. The clustering seems to improve the greater number of clusters we introduce. This will mean that the distance between members of clusters is always on average significantly bigger than the distance to the rest of the clusters. This could be explained by the fact that the coefficients of Leontief explain such a complex economy that the correlations between some industries may be overshadowed by the massive differences with other industries in the same clusters. This second result on optimal number of clusters leaves no need to further analyse the optimal solution. What may be interesting to analyse in deepness is those industries that are left clustered when the number of clusters is close to maximum.

### 4.4.2. SQUARED EUCLIDIAN DISTANCE:

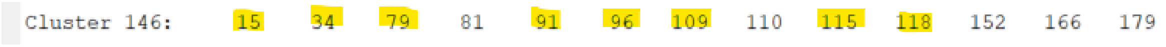
As previously explained once a centroid is in a position all points are associated to the closest centroid. The differences that may arise is how to measure the distance to that centroid, that wants to be minimised, since it is not a multidimensional problem. In this method the distance between a point (xi) and a centroid (ci) is calculated for each dimension (p), to later be squared and added:

$$d(x,c) = \sum_{i=1}^{p} (x_i - c_i)^2$$

*Equation 11: squared Euclidean distance.*

As it would be expected this method is great for creating clusters based on the sum of squared differences. The problem is that only relevant results will arise when carrying out the analysis for at least 180 centroids. We can pinpoint the points that remain together despite the large distances between objects.

For k=180:



```
Cluster 146:    15   34   79   81   91   96   109   110   115   118   152   166   179
```

*Figure 13: Composition of composite cluster, with highlighted repetitions.*

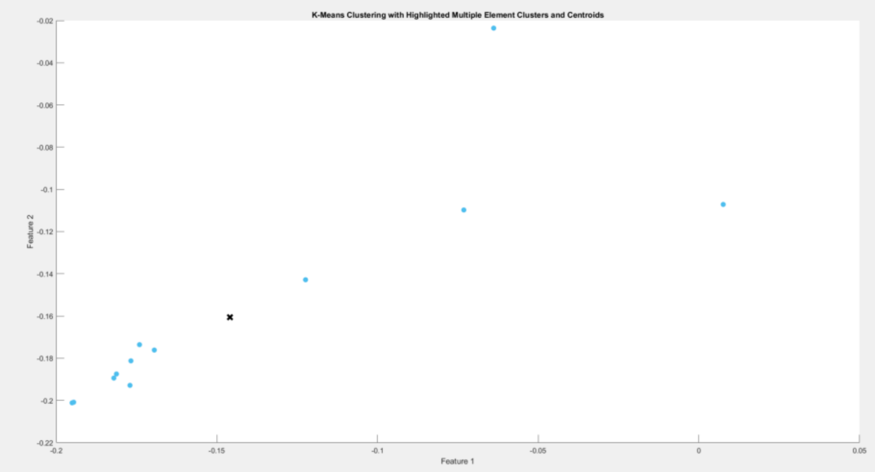


*Figure 14: K-means clustering on the economy for k=180*

*The problem that arises is that depending on the initial random distribution of the centroids the clustering results may change; this could be because industries have distinct interactions or to a large amount of dimensional noise. Results obtain lead towards 179 individual clusters and 1 cluster containing the rest of the elements. Even though with each iteration the composition of the big cluster varies, there seems to be some element that appears with each result these are Construction (15), Petroleum and coal products manufacturing (34), Motor vehicle manufacturing (79), Wholesale trade (91), Air transportation (96), Wired telecommunications carriers (109), Securities, commodity contracts, investments, and funds and trusts (115) and Real State (118).*

Some arguments may explain the relations from different points of view. Such as sector similarity between Construction (15) and Real State (118), due to the symbiotic relation they share. The first acts as the backbone of the second since Real State depends on the quality techniques and development of the construction industry. From the other's perspective, it could be considered that Real State acts as a catalyst for Construction since the demand changes in the first industry lead to developments in the other [21]. Supply chain connections such as air transportation being crucial for petroleum and motor vehicle materials being imported and exported. Communications should be the backbone of the economy, Wired Telecommunications Carriers (109), but it seems it only affects strongly enough the demand for the previously highlighted industries. Economically it could imply that it is a resilient one since the highlighted industries do not compose the same sector, some are goods and services, financial or infrastructure related, but they seem to have a stronger relation than others.

The relations between these industries seem to be clear on a bilateral basis, but it is extremely hard to find a common nexus between all these industries due to the complexity of the picture. All we know is that they seem to show a closer behaviour towards changes in demand than all the other industries. K-means has proven that the differences between industries seem to be too varied to be clustered, they show little interdependence, but we can appreciate that it has highlighted the group with more interconnections between industries. The fact that these constantly appear shows strong similarities in their Leontief coefficients and therefore in their reactions to changes in demand changes. This should be considered when taking economic measures. It will be also interesting to analyse which are the first industries to separate from the main cluster, which would be the most independent and less dependent on the rest of the economy. Being k=3 we expect to obtain the two industries that meet the previous characteristics although they do not seem to be enough to explain the economy.

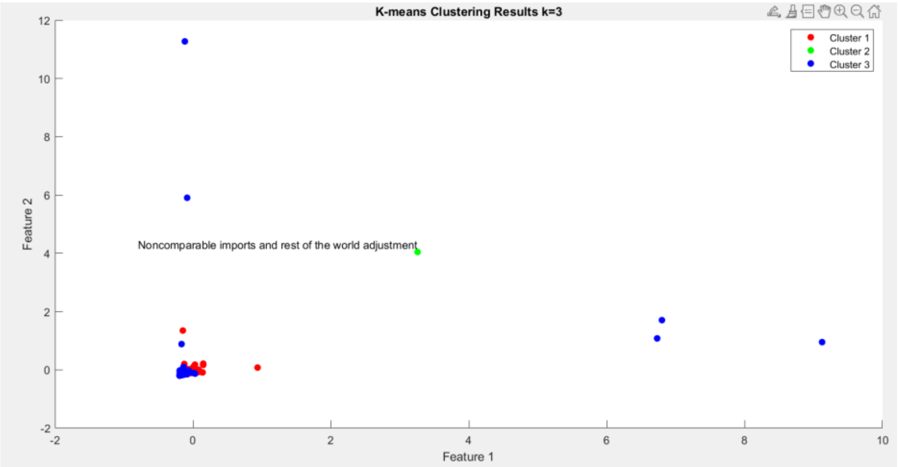


*Figure 15: K-means clustering on the economy for k=180*

Similar to the previous k-means analysis, results need to be validated, since depending on the iterations the result varies. Clusters 1 and 3 may vary their compositions, but Cluster 2 is systematically composed by Noncomparable imports and rest of the world adjustment

(192) solely. The fact that the composition of some clusters varies indicates that the relations between the changes in output is flexible and that have adaptable relationships. This internal variability could indicate that economic output may depend on external factors to this matrix. These could be factors such as policy or monetary changes. The fact that industry 192 is systematically isolated, would mean a limited influence by changes in demand on other sectors, which is surprising since some imported materials may be strictly demanded by other industries. This could be explained by having a reacting activity rather than a proactive one. Not forgetting that this industry represented the major variability of the economy, which would mean that the economy is more dependent on the changes on demand of the industry rather than the other way around.

# CHAPTER 5: RESULTS AND CONCLUSION

## *5.1 PRINCIPAL COMPONENT ANALYSIS (PCA)*

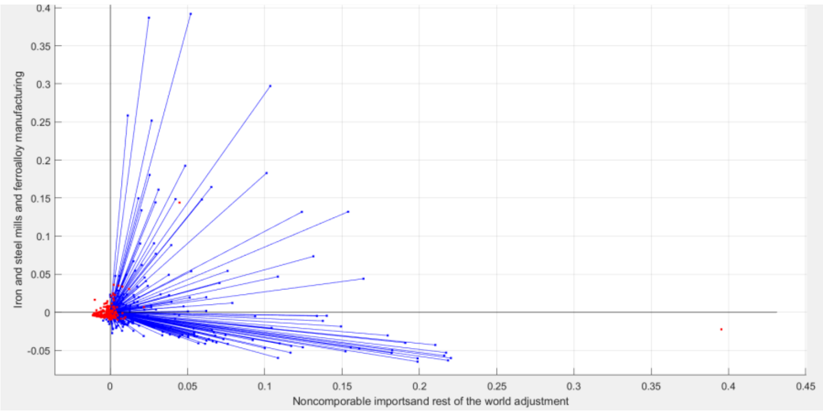


*Figure 2 (repeated): PCA sample.*

Through the principal component analysis carried out on the whole database it was revealed Noncomparable imports and rest of the world adjustment is the most dominant variable, explaining 10.27% of the total variance. Later it was discovered that this representation of the variance was no easy task, which is strongly explained by the heavy dependence on the industry by the economy. Apart from this relevant result, it was difficult to discern any other information in this complex economy, this is why it was crucial to carry out the sectorial PCA.

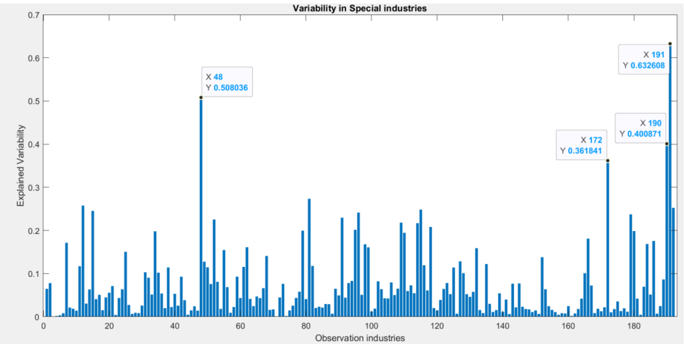
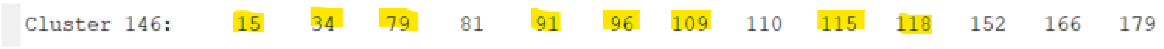


*Figure 9 (repeated): PCA on SPECIAL INDUSTRIES*

Finance and Health care sectors seem to be independent due to the little effects changes in demands of other sectors have on the output of the sector. Agriculture, forestry, fishing and hunting; and mining seem to follow a similar pattern, with the difference that there was an industry from outside the sector that affected significantly the output. The most chaotic partial PCA was of the special industries, which seem to be influenced by many other industries' changes in demand. Curiously Noncomparable imports and rest of the world adjustment which explained the major variability of the economy did not represent the major peak in the partial analysis of its sector, meaning that its influence is spread around many other sectors as seen in the figure. This is why it may have been a good representative of principal component in such a complex economy.

## 5.2 K-MEANS ANALYSIS

K-means was used to find out how the industries partition based on the changes of output caused by shifts in demand, which could not be observed through PCA alone. As previously mentioned, the first objective was to find out the optimal number of cluster, that showed the individuality of sectors and industries, without relying on a redundant result of all industries forming their own clusters. Funny enough, this was the exact result obtained, both the Elbow method and Silhouette analysis showed that the complexity and diversity of the economic data would not allow for analysis with less than 180 clusters from a database of 192 industries. The cluster composition varied through iterations, but there is a persistent presence of some industries, previously highlighted, such as Construction and Real State which may indicate a stronger interrelation that stands out in the big database, and that should be considered in policy making.



```
Cluster 146:    15   34   79   81   91   96  109  110  115  118  152  166  179
```

*Figure 12 (repeated): Composition of composite cluster, with highlighted repetitions.*

## 5.3 *SUMMARY AND FINDINGS*

The PCA analysis has helped to identify some key variables that affect the American economy's variance, in particular Noncomparable imports and rest of the world adjustment. Not forgetting individual sectors PCA analysis, which proved to be more revealing on critical interdependencies and influential industries. An example of this would be the dependence of the Steel industry on recycled materials and the high dependence of Agriculture industry on support activities. While the complexity of the database was a challenge, certain nuggets of information emerged which would help understand certain areas.

K-means analysis really shows the complexity of the relations in changes in output. The fact that the optimal number of clusters was such a big amount meant that each industry represented a similar piece of the variability of the economy. It highlighted the individuality of most industries, which was later restated by the changes in members of the remaining cluster based on the iterations. Nevertheless, some industries remained clustered despite iterations which indicates there are some focal points for economic policies, such as construction and real state being symbiotic in their outputs. The highlighted industries do not compose the same sector of the economy so it could imply stronger symbiotic relations one to one which stand out in such a complex economy since it is hard to find a common nexus.

## 5.4 *IMPLICATIONS ON ECONOMIC POLICY*

Given the complexity revealed by the analysis, an approach that considers the high interdependence of the industries of the same sector would be essential for policy making. For example, by introducing incentives for increasing Insurance Carriers and Related Activities (116) it could also affect significantly the housing sector. In a similar way, some industries seem to have something close to a symbiotic relationship, such as construction and real estate, so policies flowing in either direction could lead to a ripple effect on the counterpart, which is a significant benefit.

## *5.5 LIMITATIONS AND CHALLENGES*

The reliance on real values to avoid price distortion presents some limitations, as may undermine some variability of the economy. Such as the different effects may have on the industries, but since the analysis is carried out in one year and is already complex by itself, it is a justified cost.

While the elbow method and silhouette score are great ways to find out the optimal number of clusters of a sample, they are not perfect. In fact, obtaining different results from both processes is proof of the complexity and variability the sample contains. Which was later pictured in the different composition of clusters based on the iteration of the process, making results less consistent and reliable. Meaning we have yet to learn a lot about the economy.

Machine learning is a powerful tool, but it also has limitations. On the one hand, PCA reduces the dimensionality of the data to two dimensions, which normally results in a loss of information, especially with the variability of the matrix being so distributed, allowing interdependencies to be lost. On the other hand, K-means analysis focuses on grouping data points which ignores potential connections between sectors that could have appeared.

Although the objective was to focus on changes due to demand shifts, from the data point of view, one could argue that the economy is looked through a scope and that many other variables are ignored. Which could have enriched the view on the clustering allowing it to be more generalizable to other economies. Also, this hidden knowledge may be causing misinformation on the behaviour of the economy in the stated year, which could not be proved due to the extension of the analysis that it would require.

With the objective of having enough studies on the analysed year, many emerging recent industries and economic changes may have led to some limitations. Since the economy is constantly changing and the analysis was carried out on a ten-year-old information it may have some differences with the actual picture. It will require real-time data for political implementations, although it may contain some mistakes that are generally adjusted with time.

## *5.6 FUTURE RESEARCH DIRECTIONS*

When analysing this economy, it was stated repeatedly that the complexity was neglecting the possibility of obtaining clear conclusions. This makes sense when considering the American economy as the biggest and most shifting of the world. Since machine learning has yet to evolve to comprehend it fully, it would be interesting to start with a smaller and simpler one to find patterns that can later be extrapolated against the American economy.

As previously mentioned, the scope of the study may be a limitation. It could be interesting to consider in future research some variables such as inflation effects on the industry's output. This may lead to very different conclusions, but as proved by the extensive complexity, constantly mentioned, it will add a layer that only could be solved in a smaller scale data sample.

Based on the results obtained, although not fully clear, it would be interesting to test predictive models and changes in policy development. Changing from machine into a supervised approach to find out how on track the method can be and how it could become a powerful tool for future policy making or whether it can potentially be.

# CHAPTER 6. REFERENCES

[1] A. Phillips, «The Tableau Économique as a Simple Leontief Model,» *Oxford Academic*, pp. 137-144, 01 February 1955.

[2] L. Walras, «Eléments d'économie politique pure,» *Revue de Théologie et de Philosophie et Compte-Rendu Des Principales Publications Scientifiques*, p. 628–632, 1874.

[3] J. D. Sutomo, Statistik Indonesia, 2000.

[4] L. Walras, Elements D'Economie Politique, Paris: Thoerie de la Richesse Sociale, 1874.

[5] E. Dietzenbacher, «Interregional Multipliers: Looking Backward, Looking Forward,» *Regional Studies*, pp. 125-136, 2002.

[6] S. F. Andreas Freytag, «Sectoral linkages of financial services as channels of economic development—An input–output analysis of the Nigerian and Kenyan economies,» *Review of Development Finance*, vol. 7, nº 1, pp. 36-44, 2017.

[7] G. James, D. Witten, T. Hastie y R. Tibshirani, An Introduction to Statistical Learning, Los Angeles: Springer, 2014.

[8] M. Cave, «Wassily Leontief: Input—Output and Economic Planning,» Palgrave Macmillan, London, 1981.

[9] F. Sancho, «An Armington–Leontief model,» *Economic Structures*, 2019.

[10] L. R. Teigeiro, «Rasgos estructurales de la econmía andaaluza.,» Consejería de Economía y Conocimiento, Junta de Andalucia, 2018.

[11] D. A. &. KOLOKONTES, A. KONTOGEORGOS, E. Loizou y F. CHATZITHEODORIDIS, «KEY-SECTORS ATTRACTIVENESS OF THE GREEK ECONOMY: AN,» de *Euro-American Association of Economic Development*, 2018.

[12] Office of Financial Management of the Washington State, «Office of finantial Management,» 11 February 2021. [En línea]. Available: https://ofm.wa.gov/washington-data-research/economy-and-labor-force/washington-input-output-model/2012-washington-input-output-model. [Último acceso: 2024].

[13] G. CIUBOTARIU y L. . M. CRIVEI, «ANALYSING THE ACADEMIC PERFORMANCE OFSTUDENTS USING UNSUPERVISED DATA MINING,» STUDIA UNIV. , BABES¸–BOLYAI, INFORMATICA, 2019.

[14] H. L. C. G. B. &. M. W. Wang, Interviewee, *Does AI-based credit scoring improve financial inclusion? Evidence from online payday lending.*. [Entrevista]. 1984.

[15] A. S. Hall, «Machine Learning Approaches to Macroeconomic Forecasting,» 2018.

[16] S. D. N. K. R. K. A. Anoop Kumar, «Unveiling the Impact of Macroeconomic Policies: A Double Machine Learning Approach to Analyzing Interest Rate Effects on Financial Markets,» *Cornell University*, 2024.

[17] K. Argyrios D., K. Achilleas, L. Efstratios y . C. Fotios, «Input-Output Models and Derived Indicators: A Critical Review,» *Sciendo, Scientific Annals of Economics and Business*, vol. 3, nº 66, p. 270, 2019.

[18] Matlab, «Mathworks, Help Center,» [En línea]. Available: https://www.mathworks.com/help/stats/pca.html. [Último acceso: 2024].

[19] Matlab, «Mathworks, Help Center,» [En línea]. Available: https://www.mathworks.com/help/stats/kmeans.html. [Último acceso: 2024].

[20] National Marine Fisheries Service, «2020 FISHERIES,» 2022.

[21] Z. Rushirajsinh, «The Elbow Method: Finding the Optimal Number of Clusters,» Medium, 4 Nov 2023. [En línea]. Available: https://medium.com/@zalarushirajsinh07/the-elbow-method-finding-the-optimal-number-of-clusters-d297f5aeb189.

[22] O. G. López, «Entendiendo el Within Cluster Sum of Squares (WCSS),» Medium, 2 Feb 2024. [En línea]. Available: https://medium.com/@oriolgilabertlopez/entendiendo-el-within-cluster-sum-of-squares-wcss-14935cb64672.

[23] M. Seo, «Karma Construction Group,» 17 july 2023. [En línea]. Available: https://www.karmaconstructiongroup.com/post/the-symbiotic-relationship-between-real-estate-and-construction-a-comprehensive-analysis.

[24] F. Murtagh y P. Contreras, «Algorithms for hierarchical clustering: an overview.,» *WIREs Data Mining Knowl Discov,* vol. 2, pp. 86-87, 2012.

[25] Matlab, «Mathworks Help Center,» [En línea]. Available: https://www.mathworks.com/help/stats/hierarchical-clustering.html. [Último acceso: 2024].

[26] Esri, «ArcGis Pro toolreference,» [En línea]. Available: https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/how-density-based-clustering-works.htm#.

[27] Matlab, «Mathworks Help Center,» [En línea]. Available: https://www.mathworks.com/help/stats/dbscan.html. [Último acceso: 2024].

# ANNEX I: LIST OF ANALYZED INDUSTRIES

| | |
|---|---|
| | **Agriculture, forestry, fishing and hunting** |
| 1 | Crop production |
| 2 | Animal production and aquaculture |
| 3 | Forestry |
| 4 | Logging |
| 5 | Fishing, hunting and trapping |
| 6 | Support activities for agriculture and forestry |
| | **Mining** |
| 7 | Oil and gas extraction |
| 8 | Coal mining |
| 9 | Metal ore mining |
| 10 | Nonmetallic mineral mining and quarrying |
| 11 | Support activities for mining |
| | **Utilities** |
| 12 | Electric power generation, transmission and distribution |
| 13 | Natural gas distribution |
| 14 | Water, sewage and other systems |
| | **Construction** |
| 15 | Construction |
| | **Manufacturing** |
| 16 | Animal food manufacturing |
| 17 | Grain and oilseed milling |
| 18 | Sugar and confectionery product manufacturing |
| 19 | Fruit and vegetable preserving and specialty food manufacturing |
| 20 | Dairy product manufacturing |
| 21 | Animal slaughtering and processing |
| 22 | Seafood product preparation and packaging |
| 23 | Bakeries and tortilla manufacturing |
| 24 | Other food manufacturing |
| 25 | Beverage and tobacco manufacturing |
| 26 | Textile mills and textile product mills |
| 27 | Apparel, leather and allied product manufacturing |
| 28 | Sawmills and wood preservation |
| 29 | Veneer, plywood, and engineered wood product manufacturing |
| 30 | Other wood product manufacturing |
| 31 | Pulp, paper, and paperboard mills |
| 32 | Converted paper product manufacturing |
| 33 | Printing and related support activities |
| 34 | Petroleum and coal products manufacturing |
| 35 | Basic chemical manufacturing |

| | |
|---|---|
| 36 | Resin, synthetic rubber, and artificial synthetic fibers and filaments manufacturing |
| 37 | Pesticide, fertilizer, and other agricultural chemical manufacturing |
| 38 | Pharmaceutical and medicine manufacturing |
| 39 | Paint, coating, and adhesive manufacturing |
| 40 | Soap, cleaning compound, and toilet preparation manufacturing |
| 41 | Other chemical product and preparation manufacturing |
| 42 | Plastics product manufacturing |
| 43 | Rubber product manufacturing |
| 44 | Clay product and refractory manufacturing |
| 45 | Glass and glass product manufacturing |
| 46 | Cement and concrete product manufacturing |
| 47 | Lime, gypsum and other nonmetallic mineral product manufacturing |
| 48 | Iron and steel mills and ferroalloy manufacturing |
| 49 | Steel product manufacturing from purchased steel |
| 50 | Alumina and aluminum production and processing |
| 51 | Nonferrous metal (except aluminum) production and processing |
| 52 | Foundries |
| 53 | Forging and stamping |
| 54 | Cutlery and handtool manufacturing |
| 55 | Architectural and structural metals manufacturing |
| 56 | Boiler, tank, and shipping container manufacturing |
| 57 | Hardware manufacturing |
| 58 | Spring and wire product manufacturing |
| 59 | Machine shops; turned product; and screw, nut, and bolt manufacturing |
| 60 | Coating, engraving, heat treating, and allied activities |
| 61 | Other fabricated metal product manufacturing |
| 62 | Agriculture, construction, and mining machinery manufacturing |
| 63 | Industrial machinery manufacturing |
| 64 | Commercial and service industry machinery manufacturing |
| 65 | Ventilation, heating, air-conditioning, and commercial refrigeration equipment manufacturing |
| 66 | Metalworking machinery manufacturing |
| 67 | Engine, turbine, and power transmission equipment manufacturing |
| 68 | Other general purpose machinery manufacturing |
| 69 | Computer and peripheral equipment manufacturing, excluding digital camera manufacturing |
| 70 | Communications equipment manufacturing |
| 71 | Audio and video equipment manufacturing |
| 72 | Semiconductor and other electronic component manufacturing |
| 73 | Navigational, measuring, electromedical, and control instruments manufacturing |
| 74 | Manufacturing and reproducing magnetic and optical media |
| 75 | Electric lighting equipment manufacturing |
| 76 | Household appliance manufacturing |
| 77 | Electrical equipment manufacturing |
| 78 | Other electrical equipment and component manufacturing |
| 79 | Motor vehicle manufacturing |
| 80 | Motor vehicle body and trailer manufacturing |
| 81 | Motor vehicle parts manufacturing |
| 82 | Aerospace product and parts manufacturing |
| 83 | Railroad rolling stock manufacturing |

| | |
|---|---|
| 84 | Ship and boat building |
| 85 | Other transportation equipment manufacturing |
| 86 | Household and institutional furniture and kitchen cabinet manufacturing |
| 87 | Office furniture (including fixtures) manufacturing |
| 88 | Other furniture related product manufacturing |
| 89 | Medical equipment and supplies manufacturing |
| 90 | Other miscellaneous manufacturing |
| | **Wholesale trade** |
| 91 | Wholesale trade |
| | **Retail trade** |
| 92 | Motor vehicle and parts dealers |
| 93 | Food and beverage retailers |
| 94 | General Merchandise retailers |
| 95 | All other retail |
| | **Transportation and warehousing** |
| 96 | Air transportation |
| 97 | Rail transportation |
| 98 | Water transportation |
| 99 | Truck transportation |
| 100 | Transit and ground passenger transportation |
| 101 | Pipeline transportation |
| 102 | Scenic and sightseeing transportation and support activities for transportation |
| 103 | Couriers and messengers |
| 104 | Warehousing and storage |
| | **Information** |
| 105 | Newspaper, periodical, book, and directory publishers[1] |
| 106 | Software publishers |
| 107 | Motion picture and sound recording industries |
| 108 | Radio and television broadcasting, media streaming distribution services, social networks, and other media networks and content providers[1] |
| 109 | Wired telecommunications carriers |
| 110 | Wireless telecommunications carriers (except satellite) |
| 111 | Satellite, telecommunications resellers, and all other telecommunications |
| 112 | Computing infrastructure providers, data processing, web hosting, and related service |
| 113 | Web search portals, libraries, archives, and other information services[1] |
| | **Finance and insurance** |
| 114 | Monetary authorities - central bank, credit intermediation, and related activities |
| 115 | Securities, commodity contracts, investments, and funds and trusts |
| 116 | Insurance carriers |
| 117 | Agencies, brokerages, and other insurance related activities |
| | **Real estate and rental and leasing** |
| 118 | Real estate |

| | |
|---|---|
| 119 | Automotive equipment rental and leasing |
| 120 | Consumer goods rental and general rental centers |
| 121 | Commercial and industrial machinery and equipment rental and leasing |
| 122 | Lessors of nonfinancial intangible assets (except copyrighted works) |
| | **Professional, scientific, and technical services** |
| 123 | Legal services |
| 124 | Accounting, tax preparation, bookkeeping, and payroll services |
| 125 | Architectural, engineering, and related services |
| 126 | Specialized design services |
| 127 | Computer systems design and related services |
| 128 | Management, scientific, and technical consulting services |
| 129 | Scientific research and development services |
| 130 | Advertising, public relations, and related services |
| 131 | Other professional, scientific, and technical services |
| | **Management of companies and enterprises** |
| 132 | Management of companies and enterprises |
| | **Administrative and support and waste management and remediation services** |
| 133 | Office administrative services |
| 134 | Facilities support services |
| 135 | Employment services |
| 136 | Business support services |
| 137 | Travel arrangement and reservation services |
| 138 | Investigation and security services |
| 139 | Services to buildings and dwellings |
| 140 | Other support services |
| 141 | Waste management and remediation services |
| | **Educational services** |
| 142 | Elementary and secondary schools; private |
| 143 | Junior colleges, colleges, universities, and professional schools; private |
| 144 | Other educational services; private |
| | **Health care and social assistance** |
| 145 | Offices of physicians |
| 146 | Offices of dentists |
| 147 | Offices of other health practitioners |
| 148 | Outpatient care centers |
| 149 | Medical and diagnostic laboratories |
| 150 | Home health care services |
| 151 | Other ambulatory health care services |
| 152 | Hospitals |
| 153 | Nursing and residential care facilities |
| 154 | Individual and family services |
| 155 | Community food and housing, emergency and other relief services, and vocational rehabilitation services |
| 156 | Child day care services |

| | |
|---|---|
| | **Arts, entertainment, and recreation** |
| 157 | Performing arts companies |
| 158 | Spectator sports |
| 159 | Arts and sports promoters and agents and managers for public figures |
| 160 | Independent artists, writers, and performers |
| 161 | Museums, historical sites, and similar institutions |
| 162 | Amusement parks and arcades |
| 163 | Gambling industries (except casino hotels) |
| 164 | Other amusement and recreation industries |
| | **Accommodation and food services** |
| 165 | Accommodation |
| 166 | Food services and drinking places |
| | **Other services (except public administration)** |
| 167 | Automotive repair and maintenance |
| 168 | Electronic and precision equipment repair and maintenance |
| 169 | Commercial and industrial machinery and equipment (except automotive and electronic) repair and maintenance |
| 170 | Personal and household goods repair and maintenance |
| 171 | Personal care services |
| 172 | Death care services |
| 173 | Drycleaning and laundry services |
| 174 | Other personal services |
| 175 | Religious organizations |
| 176 | Grantmaking and giving services and social advocacy organizations |
| 177 | Civic, social, professional, and similar organizations |
| 178 | Private households |
| | **Government** |
| 179 | Federal general government defense |
| 180 | Federal general government nondefense |
| 181 | Postal Service |
| 182 | Federal electric utilities |
| 183 | Other federal government enterprises |
| 184 | State and local government educational services |
| 185 | State and local government hospitals and health services |
| 186 | State and local government other services |
| 187 | State and local government passenger transit |
| 188 | State and local government electric utilities |
| 189 | Other state and local government enterprises |
| | **Special industries** |
| 190 | Owner-occupied dwellings |
| 191 | Scrap, used and secondhand goods |
| 192 | Noncomparable imports and rest of the world adjustment |

# ANNEX II: MATLAB CODE

## Extraction of data and Leontief matrix build up

```
clear;

input = 'USE';

sectorIC = 192;
sectorFD = 132;

fileName = 'SectorPlan30.xlsx';
sheetName = 'Stubs';
startRow = 1;
numRows = sectorIC;
startCol = 1;
numCols = 4;

endRow = startRow + numRows;
endCol = startCol + numCols - 1;

startColLetter = num2col(startCol);
endColLetter = num2col(endCol);

range = sprintf('%s%d:%s%d',startColLetter,startRow,endColLetter,endRow);

idxRow = readtable(fileName,'Sheet',sheetName,'Range',range);
idxRow.BLS_IO_Summary = [];
idxRow.NAICS_2022 = [];

%{
fileName = 'FDSectorPlan30.xlsx';
sheetName = 'Stubs';
startRow = 1;
numRows = sectorFD;
startCol = 1;
numCols = 2;
endRow = startRow + numRows;
endCol = startCol + numCols - 1;
startColLetter = num2col(startCol);
endColLetter = num2col(endCol);
range = sprintf('%s%d:%s%d',startColLetter,startRow,endColLetter,endRow);
idxCol = readtable(fileName,'Sheet',sheetName,'Range',range);
%}

fileName = 'REAL_USE.xlsx';
numCols = sectorIC;
idxCol = idxRow;
```

```
year = 2012;
sheetName = num2str(year);
startRow = 2;
startCol = 2;
numRows = sectorIC;

endRow = startRow + numRows - 1;
endCol = startCol + numCols - 1;

startColLetter = num2col(startCol);
endColLetter = num2col(endCol);

range = sprintf('%s%d:%s%d',startColLetter,startRow,endColLetter,endRow);

dataTable = readtable(fileName,'Sheet',sheetName,'Range',range,...
    'ReadVariableNames',false);

fdCol = readtable(fileName,'Sheet',sheetName,'Range','GL2:GL193',...
    'ReadVariableNames',false);
fdCol = table2array(fdCol);

data = table2array(dataTable);

gpCol = sum(data,2)+fdCol;      % gross production
gpCol(gpCol==0,1) = 1e-5;       % correction for regularization

% Leontief
data = data./gpCol;
I = eye(size(gpCol,1));
data = inv(I-data);
data=normalize(data); % revisar cuando nnormalizar
```

## PCA general analysis main input and outputs.

```
[coeff,score,~,~,explained,~] = pca(data);
biplot(coeff(:,1:2),'scores',score(:,1:2));

for col=1:10
    s_col = find(coeff(:,col)==max(abs(coeff(:,col))));
    s_row = find(data(:,s_col)==max(abs(data(:,s_col))));
    msg = sprintf('output %s, whose main input %s, explains %2.2f',...
        idxCol.SectorTitle{s_col},idxRow.SectorTitle{s_row},explained(col));
    disp(msg);
end
```

### *Parcial PCA*

```
filename = 'SectorPlan30.xlsx';
sheet = 'Industry Commodity Sectors';

[~, ~, raw] = xlsread(filename, sheet);

sectorNumbers = raw(5:end, 1); %  from row 5
descriptions = raw(5:end, 2);

% Initialize variables
currentAggregation = '';
aggregations = {};

% up to sector 192
for i = 1:length(descriptions)
    description = descriptions{i};
    sectorNumber = sectorNumbers{i};

    if isnan(sectorNumber)
        currentAggregation = description;
    else
        if ~isempty(currentAggregation) & sectorNumber <= 192
            aggregations = [aggregations; {sectorNumber, currentAggregation}];
        end
    end
end

% Convert to table
aggregations = cell2table(aggregations, 'VariableNames', {'SectorNumber',
'Aggregation'});
aggregations = aggregations(1:192, :);
disp(aggregationsTable);


% aggregation names
industryNames = idxCol{:,2};
aggregationNames = aggregations{:,2};
[uniqueAggregations, ~, aggregationIdx] = unique(aggregationNames, 'stable');

resultTable = table;

% Select the desired aggregation
selectedAgglomerationIndex = 1;

currentAggregation = uniqueAggregations{selectedAgglomerationIndex};

aggregationIndices = find(aggregationIdx == selectedAgglomerationIndex);

aggregationData = data(aggregationIndices, :);

% PCA
[coeff, score, ~, ~, explained] = pca(aggregationData);
```

```
totalExplained = cumsum(explained);
numComponents = find(totalExplained >= 95, 1);
variabilityExplained = sum(abs(coeff(:, 1:numComponents)), 2);

% bar chart
figure;
bar(variabilityExplained);
title(['Variability in ' currentAggregation]);
xlabel('Observation industries');
ylabel('Explained Variability');
xticks(20:20:192);
```

## Elbow method and silhouette score

```
min_clusters = 2;
max_clusters = 191;

% Initialize
wcss = zeros(max_clusters - min_clusters + 1, 1);
silhouette_avg = zeros(max_clusters - min_clusters + 1, 1);

for k = min_clusters:max_clusters

    [idx, C, sumd] = kmeans(data, k, 'Replicates', 10);
     wcss(k - min_clusters + 1) = sum(sumd);
    silhouette_vals = silhouette(data, idx);
    silhouette_avg(k - min_clusters + 1) = mean(silhouette_vals);

    % Show progress
    fprintf('K-means clustering done for k = %d\n', k);
end

% Elbow method
figure;
plot(min_clusters:max_clusters, wcss, '-o');
xlabel('Number of clusters (k)');
ylabel('WCSS');
title('Elbow method for optimal number of clusters');
grid on;

% Silhouette Index
figure;
plot(min_clusters:max_clusters, silhouette_avg, '-o');
xlabel('Number of clusters (k)');
ylabel('Average silhouette score of objects');
title('Silhouette analysis for optimal number of clusters');
grid on;
```

### K-means for k=180

```
% Number of clusters
k = 180;

% Perform k-means clustering with 'sqeuclidean' distance, repeating 10 times
[idx, C] = kmeans(data, k, 'Distance', 'sqeuclidean', 'Replicates', 10);

multiple_element_clusters = {};
cluster_indices = [];

% Iterate through each cluster
for i = 1:k
    cluster_elements = find(idx == i);
    if length(cluster_elements) > 1
        multiple_element_clusters{end+1} = cluster_elements;
        cluster_indices = [cluster_indices; i];
    end
end

%clusters with more than one element
disp('Clusters with more than one element:');
for i = 1:length(multiple_element_clusters)
    fprintf('Cluster %d: ', cluster_indices(i));
    disp(multiple_element_clusters{i}');
end

figure;
hold on;

colors = lines(k); % Generate distinct colors for each cluster
for i = 1:length(multiple_element_clusters)
    elements = multiple_element_clusters{i};
    scatter(data(elements, 1), data(elements, 2), 50,
colors(cluster_indices(i), :), 'filled');
end

% centroids of these clusters
for i = 1:length(cluster_indices)
    plot(C(cluster_indices(i), 1), C(cluster_indices(i), 2), 'kx',
'MarkerSize', 10, 'LineWidth', 3);
end

title('K-Means Clustering with Highlighted Multiple Element Clusters and
Centroids');
xlabel('Feature 1');
ylabel('Feature 2');
hold off;
```

### K-means for k=3

```
k = 3;

[idx, C] = kmeans(data, k, 'Distance', 'sqeuclidean', 'Replicates', 10);

% store the size of each cluster
cluster_sizes = zeros(k, 1);

for i = 1:k
    cluster_sizes(i) = sum(idx == i);
end

% two smallest clusters
[~, sorted_indices] = sort(cluster_sizes);
smallest_cluster_indices = sorted_indices(1:2);
disp('Elements in the two smallest clusters:');
for i = 1:length(smallest_cluster_indices)
    cluster_elements = find(idx == smallest_cluster_indices(i));
    fprintf('Cluster %d: ', smallest_cluster_indices(i));
    disp(cluster_elements');
end

figure;
hold on;

colors = lines(k);
for i = 1:length(smallest_cluster_indices)
    elements = find(idx == smallest_cluster_indices(i));
    scatter(data(elements, 1), data(elements, 2), 50,
colors(smallest_cluster_indices(i), :), 'filled');
end

for i = 1:length(smallest_cluster_indices)
    plot(C(smallest_cluster_indices(i), 1), C(smallest_cluster_indices(i), 2),
'kx', 'MarkerSize', 10, 'LineWidth', 3);
end

title('K-Means Clustering with Highlighted Smallest Clusters and Centroids');
xlabel('Feature 1');
ylabel('Feature 2');
hold off;
```