

Master’s degree in computational social science  
2023-2024

*Master Thesis*

“Kluster It! - Integrating K-Means Clustering and Club Convergence Analysis with Shiny”

Bernard Banyuls

Carla Fernández

Madrid, 2024

**AVOID PLAGIARISM**

The University uses the Turnitin Feedback Studio program within the Aula Global for the delivery of student work. This program compares the originality of the work delivered by each student with millions of electronic resources and detects those parts of the text that are copied and pasted. Plagiarizing in a TFM is considered a **Serious Misconduct**, and may result in permanent expulsion from the University.

   
*[Include this code in case you want your Master Thesis published in Open Access University Repository]*

This work is licensed under Creative Commons **Attribution – Non Commercial – Non Derivatives**

# ABSTRACT

This thesis presents the development and implementation of a Shiny application designed to facilitate k-means clustering, club convergence analysis, and a novel Klub Index methodology. The application aims to provide researchers and data analysts with an accessible tool for performing these complex analyses with ease. The Shiny application integrates multiple R packages to perform data preprocessing, clustering, and club convergence analysis, while also offering comprehensive user instructions and visualizations to aid in interpreting the results.

## Keywords

K-Means, Club Convergence, Clustering, Shiny

# TABLE OF CONTENTS

[1. Introduction 4](#_Toc169259674)

[2. Methodology 6](#_Toc169259675)

[2.1. K-Means 6](#_Toc169259676)

[2.2. Club Convergence 7](#_Toc169259677)

[3. Shiny Application 10](#_Toc169259678)

[3.1. Objective 10](#_Toc169259679)

[3.2. How to use the ‘Kluster It!’ Shiny application 11](#_Toc169259680)

[4. Conclusions 16](#_Toc169259681)

[5. References 16](#_Toc169259682)

# Introduction

In recent years, the ability to analyze and interpret data has become one of the most important skills. Being able to understand what data may tell us is of crucial importance in terms of making informed decisions. Independently if it is in academia, industry or public policy, being able to utilize data and understand it correctly will lead to better decision-making, understanding and overall identification of patterns. Despite the potential of commonly used data techniques, they are often complex and require deep technological and statistical knowledge to implement them and understand the results as they are usually locked by some knowledge of coding and interpretation.

This complexity limits the accessibility to an audience that is not able to perform and utilize different statistical methods or programming languages. Overcoming this barrier is important to further democratize data science, allowing a wider range of users that may come from a non-technical background to work with data and utilize it. Statistical techniques play an important role not just in research but also helping organizations in different objectives. For example, a small business owner would like to understand the different types of customers they may have. This is something that clustering algorithms can do easily, but due to the technical complexity of actually performing the techniques, their accessibility is limited.

In order to address this gap, this thesis introduces a Shiny application to allow non-technical people to feel confident and comfortable utilizing clustering techniques, including K-Means clustering and Club Convergence analysis. The app will also allow incoming researchers with no technical background to introduce themselves with ease to data analysis. The main objective of the application is to integrate multiple methodologies in a user-friendly platform. It will allow users to easily perform simple clustering methods as well as a novel ‘Klub Index’ methodology that combines both K-Means clustering and Club Convergence algorithms in which the main objective is to combines multiple variables across time into a composite index. These two algorithms have the objective of clustering the data, while both providing different perspectives and objectives.

To simplify the differences between both algorithms, K-Means clustering has the objective of partitioning the data into a fixed set of groups taking into account the characteristics of the observations for different variables. On the other hand, the club convergence analysis has the objective of identifying groups of convergence among the observations by analyzing the long-term behavior or convergence patterns across the same variable or characteristic.

Both provide different insight to different objectives, for example, K-Means Clustering is useful for companies that want to segmentate their customers according to a set of characteristics, whereas the Club Convergence technique is useful in terms of analyzing long-term behavior of customers. The Club Convergence technique is mainly used in economics, primarily to analyze the possible clubs that may be formed, taking into account the behavior of countries throughout time, so that members of a group will tend to go to the same point in the future. It also provides an interesting possibility of application in other fields, for example, this technique may be applied to track and analyze the behavior of customers in terms of their purchase frequency or other customer behavior metrics.

These algorithms have been widely used on different types of research, for example, K-Means has been used by Liu et al (2020) to divide various regions of China into four clusters taking into account technology development variables for the years 2008 to 2016 and Cerqueti and Ficcadenti (2022) which they used K-Means to group several countries in terms of variables related to COVID-19 for a total of 35 countries.

Regarding the Club Convergence analysis, it is commonly seen in a variety of country-related metrics such as GDP, R&D expenditure and energy consumptions among others. For example, in terms of R&D expenditure, studies such as those by Barrios et al. (2019) and Blanco et al. (2020) have tested for convergence clubs within the European Union. Convergence in terms of GDP has been studied in different manners throughout time and the methodology has improved until the point of being able to test for the existence of convergence clubs rather than overall convergence. This advancement is largely due to the work of Phillips and Sul (2007, 2009), who developed a rigorous methodology to identify convergence clubs, thereby allowing for a more nuanced understanding of economic convergence patterns.

In the case of the K-Means clustering, this technique has already been used in the corporate sector although the Club Convergence technique is not commonly used in this sector. Despite this, there is significant potential for its application. Club Convergence could be utilized to analyze the convergence of business performance metrics, such as revenue growth, market share, or innovation rates, across different branches or subsidiaries of a corporation. By identifying convergence clubs, companies could tailor strategies to foster collaboration and knowledge sharing within these groups, ultimately driving more uniform growth and development. This would allow businesses to leverage the nuanced insights provided by Club Convergence to make more informed strategic decisions, optimize resource allocation, and enhance overall corporate performance.

# 

# Methodology

## K-Means

There are several algorithms for clustering, but the standard one is the Hartigan-Wong algorithm in which the total variance of the individuals within a cluster is defined as the sum of the squared Euclidean distances between the elements and the corresponding centroid. The centroid of each group is the center of the group that corresponds to the mean value of each individual in that cluster (Hartigan and Wong, 1979).

(1)

In this equation (1), indicates the data point i belonging to cluster and is the average value of the points in cluster .

The total variation of individuals within a cluster is defined as follows in equation (2):

(2)

Each observation is assigned to a given cluster such that the distance of the sum of squares of the observation to its centroid is minimum. The clustering algorithm follows the following processes:

* Manually define the number of clusters to use throughout the algorithm, although there are several ways to check which number k of clusters is the most optimal, which will be later explained in this section.
* The algorithm randomly places *k* centroids in the data as initial centroids. And then, each individual is assigned to the nearest centroid using the Euclidean distance
* The next step is to calculate the average value of each cluster that becomes the new centroid and the individuals are reassigned to the new centroids .
* The previous step is repeated until the centroids do not change, even if the same step is repeated again, thus achieving that the total variation of individuals within a cluster is the minimum possible. There are several ways to analyze the optimal number of centroids or clusters, such as the elbow method and the silhouette method.

Among the several algorithms to select the optimal number of cluster, the Elbow and silhouette methods are the most widely used. The elbow method uses the mean distance of the observations to their respective centroid, i.e., the total variance of the individuals within a cluster. The higher the number of clusters, the lower the variance since the maximum number of clusters is equal to the number of observations, so the number k of clusters will be the one that an increase in the number does not substantially improve the variance within a cluster.

(3)

This equation (3) is used to calculate the WCSS (Within-Cluster Sum-of-Squares) variable, which measures the variance in each cluster. The more clusters there are, the better this variable will be, to the point that it will be equal to zero when the number of clusters is equal to the number of observations.

Silhouette analysis is used to analyze the quality of clustering. It measures the separation distance between different clusters. It tells us how close each observation of a cluster is to the observations of other clusters. The range of this method whose purpose is to analyze how many clusters range from -1 to 1, and the closer the value is to 1, it means that the observation is far away from the neighboring clusters. If the coefficient is 0 it means that it is very near or on the border between the two clusters. A negative value would indicate that it is in the wrong cluster.

The silhouette method calculates the mean of the silhouette coefficients of all the observations for different values of *k*. The optimal number of clusters *k* is the one that maximizes the mean of the silhouette coefficients for a range of values of *k*.

The coefficient is calculated as follows in equation (4):

(4)

In which is the mean distance within a cluster, and is the mean distance to the observations of the nearest cluster.

## Club Convergence

For the analysis of convergence clubs, the methodology developed by Phillips and Sul (2007, 2009) is applied. This methodology allows us to study the existence of convergence clubs without having to separate our data sample into subgroups through several variables in common.

Let be expenditure in R&D. It can be decomposed as follows in equation (5):

(5)

In which would be the systematic components such as the permanent common components and would be the transitory components.

It is necessary to separate the common components from the idiosyncratic ones, so the following transformation is performed:

(6)

In which is a time-varying idiosyncratic component and is the common component. In this equation (6) captures stochastic trend behavior and measures the idiosyncratic distance between and .

Phillips and Sul (2007) propose to remove the common factor in equation (7):

(7)

Where is the variable that indicates the relative transition that measures the load coefficient in relation to the panel average at instant . That is, shows us the path of the individual relative to the panel average. Equation (8) shows us that the cross-sectional mean of is unitary and that the variance satisfies the following equation:

. (8)

Therefore, the convergence of requires that the following condition be satisfied:

(9)

Phillips and Sul (2007) have defined this condition as relative convergence and in order to specify the null hypothesis of convergence they define as:

(10)

Where can be log(t), (t) or log {log(t)}. In our case, the best choice is Phillips and Sul (2007) pose the following hypothesis test to test for the existence of absolute convergence: versus its alternative .

The following t regression model developed by Phillips and Sul (2007) is used:

for (11)

In which, if there is convergence will be 0 and, therefore will tend to infinity. For this to occur has to be greater than or equal to zero, in case it is negative the hypothesis of absolute convergence would be rejected, and we would proceed to analyze if there are convergence clubs.

For convergence clubs, Phillips and Sul (2007) developed an algorithm to identify the various clubs that might be in a sample. The following process represent how this algorithm works:

***Cross-section classification:*** The different countries are ordered in decreasing order, i.e., from highest to lowest, taking into account the values of the last period.

***Club formation***: We start by forming groups from the country with the highest value in the last period. Then we look for the first k such that when we do the log t regression test statistic, we are left with being greater than-1.65. This is done for the first two countries, and in case it is not satisfied, it is performed for the second and third countries, and so on until a pair of countries is found that does satisfy the test. In case there is no pair of countries, i.e., there is no k that meets this requirement, there would be no convergence subgroups in our data sample.

***Screening of individuals to create convergence clubs:*** In the event that in the club formation, we have encountered a pair, we proceed to perform the same test by adding countries in the order we previously classified. When the criterion is no longer met, we would have our first club.

***Recursion and stopping rule:*** A subgroup is made with the individuals that have not been screened in the previous step. The log t regression test is performed and if it is greater than -1.65, another group is formed. Otherwise, the three previous steps would be performed with this subgroup.

Schnurbus et al. (2017) would pose a fifth step, which is to merge clubs. The way it would be done would be to do the log t test regression for clubs 1 and 2, and in case it is met, we would then merge them. The same would then be done for the new club 1 and the next club, and so on until there are no more club mergers, so we would be left with the minimum number of clubs possible.

# 

# Shiny Application

## Objective

The primary objective of the Shiny application is to make advanced data analysis methodologies accessible and easy to use. Prior to that, it is important to help the user understand what each methodology does and what can they be used for. The application is mainly designed to:

1. Perform K-Means Clustering: Enable users to partition customers (or entities) into distinct clusters based on their respective characteristics in a way that customers among a group are similar to each other and different enough to other groups. This technique for example would allow us to differentiate who are the best customers, who have the potential to be one of the best among other different groups that may come from the data. For example, for the corporate sector it is common to analyze the RFM methodology which allow companies to cluster customers according to their Recency, the Frequency and the Monetary aspect. The application would allow small businesses to easily perform this clustering taking into account these three characteristics to segmentate their customers. Regarding researchers this methodology would allow to detect patterns across observations in terms of different characteristics crucial to their research of interest.
2. Conduct Club Convergence Analysis: Applying this methodology to identify groups of entities with a similar convergence behavior, meaning that, in the long run, these entities will eventually achieve similar outcomes. The idea of convergence in economics is the idea of lower income countries growing at a higher rate than high income countries implying that in the long run the lower income countries will reach the same level of economic development as their higher-income counterparts. This methodology helps understand long-term trends and dynamics within the data eventually forming groups of convergence. Although this methodology has been mainly used in academia, its potential is yet to be fully discovered.

Although both methodologies provide different objectives, it does not imply that it is not possible to combine both methodologies. The second objective of the thesis is to introduce the ‘Klub Index’ methodology that integrates the K-Means clustering and Club Convergence analysis within a single framework to leverage the strengths of both techniques. The objective is to create a composite index that would allow to understand the long-term evolution of several characteristics at the same time. The idea comes from the well-known HDI index that combines three characteristics to create an index that helps compare countries. The main goal, therefore, is to be able to create an index to analyze in a temporal manner across several characteristics.

The way the ‘Klub Index’ utilizes both methodologies is by first clustering the data using K-Means, to obtain different groups across a sample of observations in terms of different characteristics. Then, a classification model would be performed to understand which characteristics were the most important ones in terms of deciding to which cluster each observation pertains too, and using those coefficients create the composite index.

To further understand the idea, imagine a company that wants to analyze the performance of its various branches across different regions over the past decade. The company has several performance metrics such as annual revenue, customer satisfaction, marketing expenditure, employee satisfaction among others.

The first step is to perform the K-Means Clustering algorithm on the data to identify initial clusters of branches with similar performance characteristics across the years. The next step would be to calculate the average of each characteristic across time in order to perform a classification model that tries to place each branch in one of the clusters created with the main goal of identifying the importance of each characteristic in deciding which cluster each observation pertains too. In other words, how important was each characteristic in determining which cluster the observation pertains to.

The next step therefore would be for each year get the index taking into account each characteristic coefficient. In this way several characteristics have been used and collapsed into one index summarizing the overall performance of each branch each year. Then the Club Convergence analysis can be performed to identify groups of branches that follow similar trend patterns that are converging towards similar performance outcomes over time. By using the ‘Klub Index’ the company can obtain valuable insights into the key drivers of branch performance throughout time, identifying high-performing and underperforming branches, in order to develop targeted strategies to improve overall business performance. Although the example given was for an enterprise, this methodology would allow researchers to perform the club convergence analysis for countries taking different characteristics and not only one as commonly done in several studies.

## How to use the ‘Kluster It!’ Shiny application

The "Kluster it!" Shiny application is designed to facilitate advanced data analysis by making methodologies like K-Means clustering and Club Convergence accessible to users. This section provides a step-by-step guide on how to use the application effectively. Further details can be found on the repository of the application (<https://github.com/bbanyuls/KlubKluster>). A tutorial can also be found in the repository with a link to a comprehensive YouTube video guide on how to use the application.

**1. Upload Your Data**

* **Step 1**: Click on the Browse... button located at the top left of the application interface.
* **Step 2**: Select the CSV or Excel file you want to analyze. Ensure your data is properly formatted, with the first row containing the column names.

There is no need to indicate the separator, in case the csv does not use the default one as the application automatically detects if it is not the default and will ask the user to switch it. If this gave problem, there is an icon next to the upload the data button to change the separator (once the app has tried to read a csv) as well as general information.

**2. Check and Prepare Your Data**

* **Step 1**: Navigate to the General Information tab and select Check the Data Uploaded!
* **Step 2**: Verify that your data is correctly uploaded and review the data types. You can modify variable types if needed to ensure all relevant variables are numeric for analysis.

There is no need to go to check the data tab as the application will automatically move the user to that tab immediately after uploading a dataset or changing the separator which can be done either by the pop-up after the error is shown or in the information button next to the upload the data action button. It is important that the variable types of the variables the user wants to analyze to show as ‘numeric’ as the K-Means and Club Convergence algorithm are methodologies that work with distances and numeric values.

**3. Perform K-Means Clustering**

* **Step 1**: Go to the KMeans section and click on Preparation and Optimum K.
* **Step 2**: Select the variables for analysis by using the Select Variables for Analysis dropdown menu.
* **Step 3**: If necessary, check the Scale Data option to standardize your data.
* **Step 4**: Click the Run button to generate the Silhouette and Elbow method plots, which help determine the optimal number of clusters.
* **Step 5**: Review the plots to decide the optimal number of clusters, then navigate to the Analysis tab.
* **Step 6**: Enter the number of clusters in the Number of Clusters field, select the plot type, and click Run to perform the clustering.
* **Step 7**: Analyze the cluster results displayed in the table and the custom cluster plot. You can also download the cluster results using the download button.

It is important to take into account that in R, the calculation of the optimum number of clusters to be obtained may take a long period of time depending on the length of the dataset. The user may also skip this part if a specified number of groups have been previously decided due to business reasons, although it is recommended to select the number of clusters according to the metrics proposed.

**4. Conduct Club Convergence Analysis**

* **Step 1**: Navigate to the Club Convergence section and click on Data Preparation.
* **Step 2**: Reorder and rename columns as needed to prepare your data for analysis. Ensure the ID column is first, followed by year columns in ascending order.
* **Step 3**: Click on the Analysis tab within Club Convergence.
* **Step 4**: Select the range of years you want to analyze using the year slider.
* **Step 5**: Click the Run button to perform the club convergence analysis.
* **Step 6**: Choose to view all clubs or specific clubs and review the results in the provided plots or print outputs.

In this section it is important to ensure that the data used follows the indications proposed here, as the Club Convergence analysis only uses one variable, in order to allow the user to change the range of the years easily the data should be in a longitudinal manner in which the names of the variables are written in numeric. For example, if our dataset contains the GDP of European countries, the dataset should be the ID in which the actual name of the column does not matter, but the rest of the columns should follow the name of the actual year they represent. If the dataset contained months the first month should be renamed to one and follow a numerical order, this is the reason why it is provided the option to reorder the columns and rename them in case a modification should be made.

**5. Implement the Klub Index Methodology**

**Step 1: Methodology & Preparation**

* **Step 1**: Navigate to the Index Klub section and select Methodology to understand the theoretical background.
* **Step 2**: Go to Preparation to reorder and rename columns if necessary.

Due to how the ‘Klub Index’ algorithm works, in this case there is no need to reorder the columns, the user must ensure that it follows the following structure:

It is important to take into account that even if the symbol “\_” is written outside the prefix it is part of the prefix, that is to say if a column is “RD\_2008”, the prefix is “RD\_” and not “RD” in order to avoid unnecessary column names modifications if the original dataset did not have the underscore. Although the order of columns does not matter due to how the algorithm has been written, it is crucial to confirm that the prefixes are consistent across years for each variable. Additionally, ensure that the years to be indicated later are properly aligned with their respective prefixes. This consistency allows the algorithm to accurately interpret and process the data for analysis.

**Step 2: Calculate Index Coefficients**

* **Step 1**: Click on Index Coefficients.
* **Step 2**: Select the ID variable and enter the prefixes for the variables you want to analyze.
* **Step 3**: Add the prefixes using the Add Prefix button and click the play button to calculate the importance coefficients.
* **Step 4**: Review the importance plot to understand the significance of each variable.

An important thing to take into account is that the prefix needs to be common throughout the range of the variable as well as write is exactly as it would be by removing the year from the column name. For example, if the column name is “RD\_2018” the prefix written should be “RD\_”. As the user only has to write the prefixes in order to get the coefficients, in case the user makes a mistake there is a button to delete all prefixes written prior to doing the analysis. The results of the coefficients will range from 0 to 1, for example imagine we have three variables “RD”, “GDP” and “energy”, the corresponding coefficients could be 0.7, 0.2, 0.1 respectively, implying that RD was the most important variable in terms of deciding which cluster each observation pertains to in general.

**Step 3: Analyze Index Club**

* **Step 1**: Go to Index Club.
* **Step 2**: Select the ID variable, enter the prefixes, and specify the coefficients and ranking order for each variable.
* **Step 3**: Specify the start and end years for the analysis.
* **Step 4**: Click Run Index Club Analysis to generate the index and perform the club convergence analysis.
* **Step 5**: Choose to view all clubs or specific clubs and select the type of result (Plot or Print) to review the results.

In case the user does not correctly writes the prefix, the application will show that the column does not exist, for example going to the example explained before, if the user writes “RD” it would show a notification that the column “RD2018” does not exist. Although the coefficients can be calculated in the previous tab, the user may freely decide which coefficients each variable should have. The idea is that the index for each year is calculated in the following way:

(11)

Regarding the specification of the ranking order is to scale the data due to the possible differences across magnitudes due to working with several variables at the same time. Due to how the Club Convergence algorithm work we cannot scale the data in common ways such as from zero to one due to working with logarithms. In order to explain the idea of the scaling, the example of the analyzing customers in terms of their “RFM” is useful. Imagine we have a total of 2000 customers, the way it works is that the customer with the highest frequency (the number of times they have gone to the shop) would get a value of 2000 and the customer with the lowest frequency will get a value of 1.

In some cases it may be possible that we want the customer with the lowest value in a respective variable to be the “highest” for example in terms of recency that measures how many days since the last purchase, we want to positively take into account that the fewer it has been, the better, for example someone that bought something recently would have a higher value in the rank, making the rank for that respective variable, descending. In a descending ranking the largest value would get the rank of 1, the second largest value would get the rank of 2, whereas on the ascending rank the lowest value would get a rank of 1. The user may decide which type of ranking for the specific variables depending on the context.

By following these steps, users can effectively utilize the "Kluster it!" application to perform sophisticated data analyses, gain insights, and make informed decisions based on their data. The application simplifies the process, making advanced methodologies accessible even to those with limited technical backgrounds.

# Conclusions

To conclude, "Kluster it!" is a Shiny-based tool designed to simplify and enhance the process of advanced data analysis for both technical and non-technical users. By integrating K-Means clustering and Club Convergence methodologies, the application provides a comprehensive suite for analyzing complex datasets and uncovering insightful patterns and trends.

The Shiny application offers a user-friendly interface that guides users through the process of data preparation, analysis, and interpretation regarding the use of the methodologies provided which are K-Means Clustering, Club Convergence Analysis and the innovative ‘Klub Index’. By making advanced analytical methodologies accessible, the application empowers users to conduct sophisticated analyses without requiring extensive programming knowledge. The K-Means Clustering algorithm has been widely used across various sectors, whereas the Club Convergence methodology has been traditionally rooted to economic studies, one of the objectives of the thesis outside of developing the Shiny application, is to explore the integration of both techniques mentioned in an innovative methodology called ‘Klub Index’ that combines multiple variables across time into a composite index. The application ‘Kluster it!’ represents a significant step towards democratizing data analysis, providing a valuable resource with the goal of making data-driven decision-making more accessible and impactful.

Although this is a complete application, in the future changes to the application will be made in terms of adding different clustering methodologies, as well as improving the visual aspect of the instructions and steps guide of the application.

# References

A. Blanco, F., J Delgado, F., & J Presno, M. (2020). R&D expenditure in the EU: convergence or divergence? *Economic research-Ekonomska istraživanja*, *33*(1), 1685-1710. <https://doi.org/10.1080/1331677X.2020.1756371>

Barrios, C., Flores, E., & Martínez, M. Á. (2019). Club convergence in innovation activity across European regions. *Papers in Regional Science*, *98*(4), 1545-1565. <https://doi.org/10.1111/pirs.12429>

Cerqueti, R., & Ficcadenti, V. (2022). Combining rank-size and k-means for clustering countries over the COVID-19 new deaths per million. *Chaos, solitons, and fractals*, *158*, 111975. <https://doi.org/10.1016/j.chaos.2022.111975>

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *28*(1), 100–108. <https://doi.org/10.2307/2346830>

Liu, C., Gao, X., Ma, W., & Chen, X. (2020). Research on regional differences and influencing factors of green technology innovation efficiency of China’s high-tech industry. Journal of Computational and Applied Mathematics, 369, 112597. <https://doi.org/10.1016/j.cam.2019.112597>

Phillips, P. C. B., & Sul, D. (2007). Transition Modeling and Econometric Convergence Tests. *Econometrica*, *75*(6), 1771–1855. <http://www.jstor.org/stable/4502048>

Phillips, P. C. B., & Sul, D. (2009). Economic Transition and Growth. *Journal of Applied Econometrics*, *24*(7), 1153–1185. http://www.jstor.org/stable/25608787

Schnurbus, J., H. Haupt, and V. Meier. (2017). Economic transition and growth: A replication. Journal of Applied Econometrics 32: 1039–1042 <https://econpapers.repec.org/RePEc:wly:japmet:v:32:y:2017:i:5:p:1039-1042>