Methodologies and implications of clusters classification: an application to financial centres

Josselin ROMAN

September 9, 2017

Contents

1 Introduction **2**

2 Data Selection and Transformation **3**

2.1 Type of Data 3

2.2 Preparation of Data 5

2.3 Possible Forms of Classification 11

3 Methodologies **12**

3.1 Multiple-Criteria Decision Analysis 12

3.1.1 Weighted Sum 12

3.1.2 Analytic Hierarchy Process 13

3.1.3 ELECTRE Methods 16

3.2 Clustering Methods 18

3.2.1 K-Means Clustering 18

3.2.2 Hierarchical Clustering 20

3.3 Factor Analysis 23

3.4 Network Analysis 27

4 Conclusion **30**

Appendix **31**

Bibliography **34**

1 Introduction

The aim of this article is to outline existing methodologies to categorize clusters and to shed light on their respective implications. The example of financial centres (FC hereafter) will be used throughout this article, but the techniques presented should be relevant for any kind of clusters classification. Our approach will raise the following questions. Although some of them will be directly addressed, our goal is not to provide definitive answers to each of these questions but rather to stimulate debate among our readers.

- What form should a clusters ranking take? Should it be restricted to an ordinal hierarchy or show the distance between centres' ranks? Would a graph representation make sense?
- What series of data should be included in the measurement? Are national data relevant to assess a financial centre? Should we use data coming from surveys or only series coming from recognized data providers? Should we use ordinal or interval data?
- Should we rank FC according to their overall performance? Or create separated scores according to each specific activity? For a global ranking, what weight should be assigned to each activity?
- In practice, what methods do we know to aggregate heterogeneous data? What are their respective implications?

The remainder of the paper is organized as follows. In the next section, we discuss what characteristics are needed to develop a robust classification of global clusters and preliminary steps toward it. Section 3 presents different methodologies to aggregate heterogeneous variables and the last section concludes.

2 Data Selection and Transformation

In this section, we will lay out the qualities a series of data should have in order to be used in an impartial, transparent and replicable clusters ranking. On the contrary, we will show why some series should be avoided. It includes both data selection and data preparation. We will also briefly discuss the various forms this kind of classification can take.

2.1 Type of Data

The first step is to decide which series of data you want to use for your analysis. This should be done simultaneously with the choice of the methodology you will implement, as some of them require specific data. A preliminary step is also to define clearly the object you want to study. Boissinot and Marx (2017) outline the main definitions in the literature and show what drivers allow a financial centre to be successful. You will need to explain what you believe are the most important features to describe a financial cluster in today's financial system to make a relevant choice of variables.

There are usually two main types of data that are used in existing rankings: quantitative and survey. Quantitative data cannot be ignored, as they represent the only source of impartial and transparent data, which make results of the analysis more reliable. The question related to quantitative data is whether we should include private data, or fully rely on public ones? In the scope of our analysis, this is a tricky question to answer, as it is not easy to characterize a financial centre properly without resorting to private data. Thus, you should not refrain from using private data if it allows to describe better a dimension you identified to be important for FC classification. Nevertheless, it damages the replicability of the method, and public data should be utilized as long as they are available. To ensure transparency, sources of private data should also be clearly stated. Turning to data coming from surveys, the consensus is quite weaker. Although it may appear insightful to gather informations from market participants, this method also brings a lot of questions. The first question is linked to the way these data are gathered. In the *Global Financial Centres Index*[1] for instance, people are surveyed on FC they do not belong to, and their answers regarding their own financial centre are not taken into account. As it may eliminate some kind of personal bias, we can also wonder about

[1]The GFCI is considered as the reference in the field of FC rankings.

the relevance of their answers. Should this kind of classification be made on the reputation of FC, or on their intrinsic values? The second issue is related to the aggregation of preferences. On a theoretical ground, this kind of survey is subject to the impossibility theorem coined by Arrow (1951). Indeed, he showed that, as far as people are consistent in their choices, it is not possible to deduce a collective choice from individual preferences. On a more practical ground, the question is how to extract information from all these answers. Should we average them ? Keep the mode ? When mixing quantitative and data coming from surveys, what weight should be assigned to each part ? All these questions surrounding data obtained through surveys make it controversial. If one decides to use such data, he should make the survey process as well as aggregation techniques transparent.

Another distinction needs to be made between ordinal, interval and numerical data. Ordinal data refers to a natural order. Most of the time, data coming from surveys are ordinal, as people are usually asked to give a rating on a predetermined scale (from 1 to 5 for instance). Numerical data have a meaning as a measurement, such as the price per square meter or the percentage of graduates in a city. Interval data are series of numerical data that are split into several intervals (for example grouping cities where the price per square meter is somewhere between 5 000 and 6 000€). It allows to divide observations into groups while keeping a common scale. Whatever the type of data you decide to use, we will see in the next subsection that we can aggregate them to form an index, thanks to a method called normalization.

Now that you are able to decide what series of data you want to include in the analysis, we will see how to prepare it properly, before turning to the methodology part.

2.2 Preparation of Data

In multivariate analysis, it is important to transform data before using it for any of the methodologies we are going to present. We will address three main issues: criteria that need to be minimized when others should be maximized, rescaling of data and problems associated with correlation of input variables.

Inversion of variables to minimize

In *Multiple-Criteria Decision Analysis*, it often makes sense to use criteria that should be minimized. To assess the business environment in a country for instance, you may want to take into account the corporate tax rate. Unlike most criteria, you would prefer a lower value of corporate tax rate. Thus, you need to inverse values associated with this variable to make your dataset uniform.

Let a_{ij} be the performance of city i on criterion j that you initially want to minimize. Transformed data are:

$$a'_{ij} = \max_i(a_{ij}) - a_{ij}.$$

Hence, the maximum of the series become 0 and the city with the initial minimum value is transformed into the initial maximum. The other advantage of this method is that it respects the initial scale. Once you have a uniform dataset, you need to normalize it to make series comparable.

Normalization of data

Normalization (also called standardization) is a well known method in statistics that aims at rescaling a dataset to a common range, in our case to be able to aggregate data. It is often needed when some series take large values compared to others or have different measurement units. Although there are many ways to normalize data, we will only present two of them but you can find a detailed overview of normalization methodologies and their implications in Nardo et al. (2005).

The most widely used method is the *z-score standardization*. It is a powerful tool because it converts original data into series that have a 0 mean and a

standard deviation of one. These features are appreciated in statistics as it is then straightforward to run statistical tests. We compute the normalized value of an observation as follows:

$$a'_{ij} = \frac{a_{ij} - \bar{a}_j}{\sigma_j},$$

where $\bar{a}_j$ and σ_j are respectively the average and the standard deviation of values across cities i on criterion j.

The other commonly encountered method of normalization in the literature is *re-scaling*. The transformation is achieved on the range instead of the dispersion of data, so you do not need to calculate the standard deviation. The transformed series lies in the range $[0, 1]$ and we compute the normalized value of an observation as follows:

$$a'_{ij} = \frac{a_{ij} - \min\limits_i a_{ij}}{\max\limits_i a_{ij} - \min\limits_i a_{ij}}.$$

As the transformation is achieved on the range, it may imply a distortion effect if the original range is unusually large (presence of outliers) or too small. For instance, the re-scaling will induce a widening of the small range and implies a greater impact on the final composite score than it would have had using the z-score method. Furthermore, because z-score use the standard deviation as scaling factor, cities with outstanding performance on some criteria and under average on others will achieve higher ranks than cities with average performance on all criteria.

Correlation issues

The problem with correlation between variables in the context of index construction lies in the risk of overstating some dimensions of the index. Indeed, if several variables carry the same information, this particular information will be accounted for more than one time. To deal with this issue, we first need to define a way to measure correlation between variables. The literature provides several methods, we will show three of them and briefly introduce some other ones.

Pearson correlation

The most commonly used technique to compute a correlation was coined by Pearson (1895) at the end of the twentieth century. This statistic owes its popularity to the fact that it not only expresses the strength of the correlation between two variables, but also its direction. Indeed, the coefficient can take any value from -1 to 1. Most of the time, people use it without citing the author, simply calling it correlation. The underlying assumptions are that both variables are normally distributed and that they exhibit both linearity and homoscedasticity.

Consider a number of cities N and two variables x and y: x_i is the value of city i on variable x. The Pearson's correlation between these two variables is calculated as follows:

$$r = \frac{N(\sum_{i=1}^{N} x_i y_i) - (\sum_{i=1}^{N} x_i)(\sum_{i=1}^{N} y_i)}{\sqrt{[N\sum_{i=1}^{N} x_i^2 - (\sum_{i=1}^{N} x_i)^2][N\sum_{i=1}^{N} y_i^2 - (\sum_{i=1}^{N} y_i)^2]}}.$$

The main drawback of this method is that it is quite sensitive to outliers, particularly if you are using a small sample. The assumptions about the distribution of variables and their linear relation are also quite strong.

Spearman rank correlation

The Spearman rank correlation is a non-parametric statistic that measures the degree of rank correlation between two variables that was introduced in Spearman (1927). Put in other words, we are interested in the monotonicity of our variables, and not in their linear relation. Unlike Pearson's correlation, there is no particular assumption on the distribution of the variables. Nevertheless, this measure does not give any indication about the direction of the correlation as it lies between 0 and 1. A spearman correlation of 0 indicates no relationship when 1 is a perfect relationship.

To compute the Spearman rank correlation, the first thing you need to do is to rank observations. For instance if $N = 20$ and the city i has the lowest value on variable x, x_i will be equal to 20. After having ranked observations on variable y the same way, you can calculate d_i, which is the difference between city's i rank

on variables x and y. Then, the Spearman rank correlation reads as follows:

$$\rho = 1 - 6 \sum \frac{d_i^2}{N(N^2 - 1)}.$$

The advantage of this method compared to Pearson's correlation is that it is not sensitive to strong outliers, as only the rank matters, and not the absolute value. It can also be used when the series studied violate the assumptions of Pearson's correlation.

Kendall rank correlation

The Kendall rank correlation is a non-parametric measure of the strength of dependence between two variables that was proposed by Kendall (1938). As with the Spearman correlation, the Kendall correlation can take any value from 0 to 1.[2] The formula to calculate it is the following:

$$\tau = \frac{n_c - n_d}{N(N-1)/2},$$

where n_c is the number of *concordants* and n_d the number of *discordants*. To find these numbers you first need to ranks observations. Create a column where you rank the cities you want to study from 1 to N, in ascending order according to their performances on variable x. Then for each city i, report its rank y_i on variable y in the following column. The next column will contain the number of concordants, which is the number of larger ranks below a certain rank. Using the second column you can easily find n_{ci} by counting the number of ranks that are below the row i in the table and greater than y_i. n_c is equal to $\sum_{i=1}^{N} n_{ci}$. Similarly, you can find n_{di} by counting the number of ranks that are below the row i in the table and smaller than y_i and compute n_d as $\sum_{i=1}^{N} n_{di}$. This method has the same advantages as Spearman's rank correlation over Pearson's correlation, but it is a bit more complicated to implement.

Other more complicated techniques such as the *cross entropy* coming from the cryptanalysis or the *Kullback-Leibler divergence* (a measure of dissimilarity) can be used to assess the correlation between two variables. The interested reader can

[2]The statistic can provide a negative value but you will record the absolute value.

refer to Karmeshu (2003) and Kullback and Leibler (1951) for detailed explanations. In any case, it is always interesting to use several statistics to be able to compare results and validate your choices.

Once you decide which measure(s) of correlation you want to use, the first step is to edit a correlation matrix of your variables. If you detect a correlation issue (using a threshold you would have defined, for instance 0.9 for the Pearson's correlation) between some variables, you will need to treat it before any further calculation. We will now present two methods that allow to deal with this correlation issue.

The most popular tool to decorrelate variables is the *Principal Component Analysis* (PCA). The goal of this method is to summarize a large set of variables (that can be correlated) with a small number of representative variables that are independent from one another. The method is successful if these few variables are able to explain most of the variability of the original dataset. For more information about the details of this methodology, we advise readers to refer to James et al. (2013). You can perform the PCA with most of statistical software. Nevertheless, you will be faced with one major question: how many components should you keep? There are basically three different ways to answer this question. First, the easiest method is to choose a threshold y of explained variance and keep a number x of eigenvectors that allows you to reach this number. Indeed, you can consider that if x eigenvectors explain $y\%$ of the variance of the original dataset (usually at least 70%-80%), this is enough to faithfully represent the original information contained in the dataset. The second way to decide how many eigenvectors to keep is to define a threshold on eigenvalues. If an eigenvalue is superior to this threshold, you will keep the corresponding eigenvector. The *Kaiser rule* states that you should keep a component if its eigenvalue is superior to 1, which means that it contains as much information as an average single variable. Finally, you can choose the number of eigenvectors by examining the *scree plot*. The scree plot shows the number of principal components against the percentage of variance explained. Thus, you can see when adding one more component only helps marginally to explain the variance in the original dataset. This is often called the *elbow* method as the identification of the angle on the scree plot leads to the selection of the number of principal components. The problem with these techniques is that it can be quite subjective to define a threshold or to determine where the elbow is.

The other method is to look at the correlation matrix and simply drop variables that exhibit too much correlation. However, it raises the question of which variable to keep. If you find two correlated variables and don't know which one to

retain, an option is to run a PCA on these two variables and to keep appropriate components as we just explained. Then you can add these components to the remaining variables. Another approach is what we could call *recursive decorrelation*. The first step is to create groups of correlated variables. If variable x is correlated to variable y, which is itself correlated to variable z but not to variable x, these three variables can be grouped. Once again the problem is to decide which variable carry the more useful information in order to rank financial clusters. The most objective technique to answer this question is again to use the PCA on the group of variables and to include the selected components to the analysis.

2.3 Possible Forms of Classification

Depending on data you select and on the way you transform it, but also (as we will see in the next section) on the methodology you decide to use, the resulting classification can take different forms.

Firstly, you need to know what output you want to produce. With this in mind, there are typically two kind of methods: those which lead to graphical representations and those which lead to ordinal rankings. When it comes to classify FC, rankings are the most common objects you will encounter. The primary reason is that people usually prefer to be able to list cities in a hierarchical order to compare their respective importance. Nevertheless, a graph can be just as much insightful, if it is properly constructed. In this article, we will present methods resulting in rankings, and others resulting in graphical representations that can have different meanings. With network analysis, we are also going to see that it is possible to combine both of them.

Regarding rankings, another choice needs to be made. Indeed, you can either produce a hierarchy for which you only know the respective position of each city, or a ranking that shows what is the distance between cities. We usually only hear about ordinal rankings in the press, because what most people are interested in is to know how their cities, schools or countries are ranked. However, these rankings are usually based on calculations that show distance between cities and it can be interesting to find a way to take it into account when displaying results. Most of the methods we will present show gaps between clusters in addition to their rank.

Finally, for the purpose of FC analysis, it can be useful to split the global score of a city into sub-scores for each of the dimension you identified to be relevant to assess the quality of a financial cluster. In this way, you can produce one ranking for each of these dimensions. This will allow you to break down the performance of each city and determine their areas of expertise. Aggregating the different dimensions, you can get the global score that will be used to draw the final ranking. This feature seems particularly attracting, as you can divide the global score into dimensions reflecting what you value in a financial centre, but you can also use it to assess the performance of each city in different kinds of financial activities for instance. Once again, most of the methods we will lay out enable you to create these sub-rankings.

3 Methodologies

In this section, we will outline existing methodologies that can be employed to create clusters rankings. Some of them have already been used for this purpose while others have never been implemented for this kind of analysis. Having discussed each method, we will weigh arguments both for and against it in the particular case of clusters comparison. When available, references to concrete applications will also be provided. Throughout this part, we will assume that input variables have already been selected.

3.1 Multiple-Criteria Decision Analysis

3.1.1 Weighted Sum

The weighted sum method is the most commonly used to build rankings with heterogenous data. Sometimes taking as inputs both quantitative data and data coming from surveys,[3] it allows to create a composite index that can then be used to rank clusters. We will see later that besides being a method *per se*, it is also often part of other methodologies.

Methodology

Consider a number n of clusters c_i and a number m of criteria v_j. Let a_{ij} be the performance of cluster i on criterion j.

1. Depending on your preferences, assign a weight w_j to each of your input variables v_j.

2. Transform data as explained in the previous part : standardize criteria and normalize weights to 1[4].

3. Perform the weighted sum : $I(a_i) = \sum_{j=1}^{n} w_j a_{ij}, \quad \forall i \in [1, n]$.

[3] As in the case of the Global Financial Centres Index.
[4] Weights only need to be normalized if they do not sum to 1.

4. Rank clusters in decreasing order according to the value I of the index obtained.

Benefits and Drawbacks

The main advantage of this method is that it is fairly simple to implement. Once data has been prepared and weights determined, it is straightforward to find the index value of the cluster and thus its rank. It is also intuitive enough so that people who are not comfortable with statistics can understand and apply it. Being simple and easily replicable, this methodology ensures that the resultant ranking is transparent.

On the other hand, the main drawback of the weighted sum method lies in its core: the selection of weights. Indeed, the final ranking varies greatly according to the weight affected to each criteria. As it is a subjective choice, it should be clearly specified and explained to guarantee the transparency of the methodology. The result also depends on the standardization procedure used in the second step. Therefore, it can be useful to test several standardization procedures to assess the robustness of the outcome.

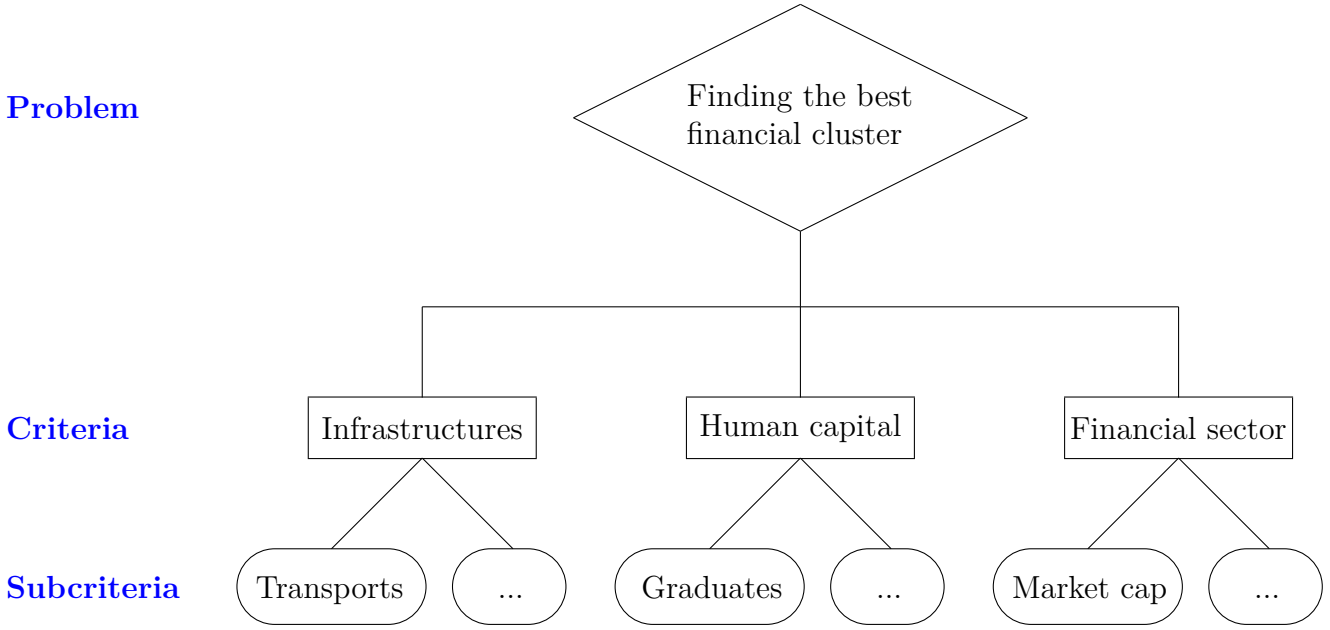
3.1.2 Analytic Hierarchy Process

Analytic Hierarchy Process is a multiple-criteria decision technique coined by Saaty (1988). It was originally designed for making group decisions but it can also be used for ranking purposes. You can find plenty of examples in the book of Saaty and Vargas (2012). As with the weighted sum method, it allows to create a composite index, which can then be utilized to rank clusters. The specificity of this methodology is that it breaks down the problem into subproblems that are solved independently before being aggregated. In the case of ranking FC, we can divide the problem into several dimensions reflecting the important qualities such clusters should have.

Methodology

1. Break down the problem (finding the "best" financial cluster) into subproblems (we will also call them criteria, for example we can compare: infrastructures,

human capital, financial sector...) and assign variables to each of the subproblems (we will also call them subcriteria, for example the quality of transports, the percentage of graduates or the market capitalization). You can find below a graphical representation of an analytic hierarchy process with 3 criteria, each one containing 2 subcriteria.



2. Compare criteria two by two and decide how important one is compared to the other. In the next step we will see how to convert this choice into numbers. In the same way, compare subcriteria that are included in each criterion two by two. The goal is to assign a weight to each criterion and each subcriterion.

3. For the final problem and for each criteria from 1 to n, write binary comparison matrices with columns i and rows j. If you believe that criteria i and j should be given the same importance, write 1 in cell (i, j) but also in cell (j, i). If you believe that criterion i is moderately more important than criterion j, write 3 in cell (j, i) and 1/3 in cell (i, j). Indeed, you compare the line criteria to the column one. Here we use the scale of Saaty and Vargas (2012), which ranges from 1 to 9 and that you can find in the appendix below.

4. Determine the weight of each criterion. First standardize columns so that criteria sum to 1: $a'_{ij} = \frac{a_{ij}}{\sum_{j=1}^{n} a_{ij}}$. Then you need to sum each line and divide the result by the number of criteria to standardize weights to unity: $w_i = \frac{\sum_{i=1}^{n} a'_{ij}}{n}$.

5. Calculate the consistency ratio: $\frac{\text{consistency index}}{\text{random consistency}}$, where the consistency index reads as follows: $\frac{\text{average consistency} - n}{n-1}$, with n being the number of criteria studied. The random consistency table can be found in appendix. The average consistency is obtained thanks to a series of calculations. First, you need to multiply each cell of the unstandardized pairwise comparison matrix by the weight affected to its line and then to sum each line. You get the consistency of each criterion by dividing the previous sum by the weight of the criterion. Taking the simple average of all criteria, you find the average consistency.
   If the consistency ratio exceeds 0.1, you need to reconsider your pairwise comparison matrix. It means that some relations you decided are inconsistent with others.

6. Repeat steps 4 and 5 for all subcriteria within each criterion.

7. Compute the real weight of each subcriterion: multiply the weight of each subcriterion by the weight of the criterion it belongs to.

8. Calculate the final index for each cluster: perform a weighted sum using the performance of the cluster on each subcriterion and the weights computed in the previous step.

9. Rank clusters in decreasing order according to the value of the index obtained.

Benefits and Drawbacks

Compared to the weighted sum method, this approach is much more complicated to implement. If you want to assess the competitiveness of a cluster on a wide range of variables, it will require either a specialized software or the knowledge of a software allowing matrix calculus (*e.g.* Matlab, Octave or R). Besides, the more criteria you include, the higher the risk of inconsistency is. You will also need to set a relative importance for each pair of subcriteria, which can be quite time consuming. Moreover it implies that all these subcriteria are comparable two by two.

The main benefit of this method is that it allows its builder to separate the problem into subproblems, which makes a lot of sense in the case of FC ranking. Indeed, you can create criteria that are useful to assess different dimensions of a cluster, and potentially create a ranking on each of these dimensions. The choice of the relative importance of each criterion is also a smoother way to find a weight than simply assigning it.

3.1.3 ELECTRE Methods

The ELECTRE methods were developed by Roy (1968), Roy and Bertier (1973) and Roy (1978) and adapted in many other articles. In this paper, we will only present ELECTRE I to give the reader an idea of the concept but more informations about the subsequent methods can be found in the bibliography. Unlike the two previous ones, this technique does not yield a hierarchical ranking, but rather a graphical representation from which you can determine a group of clusters dominating others.

Methodology

Let c represent clusters, $u_j(c_i)$ is the performance of cluster i on criterion j. If $u_j(c_i) > u_j(c_k)$, city i is said to outrank city k on criterion j.

1. Assign a weight w_j to each selected criterion j.

2. Calibration of the performance matrix: standardize each criterion to a common scale. For instance on a scale of 0 to 20, set $\min u_j(c) = 0$ and $\max u_j(c) = 20$. For all clusters in between, $\frac{u_j(c_i) - \min u_j(c)}{\max u_j(c) - \min u_j(c)} * 20$.

3. For each criteria, write a relationship matrix between cities. It is the same idea as in the AHP method, except that you will write signs instead of numbers. If $u_j(c_i) = u_j(c_k)$, write an equal sign in cell (i, k) but also in cell (k, i). If $u_j(c_i) > u_j(c_k)$, write a positive sign in cell (k, i) and a negative sign in cell (i, k).

4. Create 3 summaries matrices (one for each sign): $S^+, S^=$ and S^-. For each pairwise comparison, write for which criterion i is better/equal/worse than/to k. If $u_j(c_i) > u_j(c_k)$ for $j = \{1, 3, 4\}$, write $\{1, 3, 4\}$ in matrix S^+ cell (k, i) and inversely $\{1, 3, 4\}$ in matrix S^- cell (i, k).

5. Define the set of concordance S^C: simply write the matrix $S^+ \cup S^=$ (uniting the two matrices cell by cell). The discordance matrix S^D is equal to S^-.

6. Compute concordance indexes: for each pairwise comparison in the concordance matrix, write a matrix C. Sum the weights assigned to concerned criteria: for each cell $S^C(i,k) : C(i,k) = \sum w_j$ for all $j \in S^C(i,k)$.

7. Compute discordance indexes: for each pairwise comparison in the discordance matrix, write a matrix D:

   - $D(i,k) = 0$ if cell $S^D(i,k)$ is empty.
   - $D(i,k) = \frac{\max(u_j(c_i) - u_j(c_k))}{scale}$ with $j \in S^D$.

8. Filter clusters: choose a threshold (for instance 0.2). For each pairwise comparison, if the discordance index is under 0.2 and the concordance index is above 0.8 (1 - threshold), we say that city i outranks city j. If $C(k,i) > 0.8$ and $D(k,i) < 0.2$, then i outranks k.

9. Find the core: make a graphic representation of the previous step. Cities that are outranking others and are not being outranked themselves are part of the core. Several groups of clusters can be identified.

Benefits and Drawbacks

Depending on your ultimate goal, having a graphic representation instead of a hierarchic ranking can be seen as a benefit or a drawback. In the case of clusters classification and particularly FC, it can be interesting to perform both analyses as it gives complementary insights. The main benefit of this method is related to the robustness analysis conducted from step 5 to 8. It ensures that outranking relations that are found are strong and not subject to statistical biases. Nevertheless, this method, as the weighted sum, relies on a subjective choice of weights at the beginning of the procedure, which can lead to results subject to personal biases.

3.2 Clustering Methods

Clustering is a set of methods that were developed to find groups with common characteristics within a dataset. Unlike the ELECTRE methods, clustering won't lead to a direct outranking relation but will rather identify groups where there exists a certain degree of homogeneity between subjects. We investigate methodologies and implications of the two best-known clustering methods, which are *K-means clustering* and *Hierarchical clustering*. The following presentation is based on the book of James et al. (2013), where you can also find illustrative examples as well as codes to replicate these methods in R.

3.2.1 K-Means Clustering

K-Means clustering is a method that allows to classify N observations (FC in our case) in a predetermined number K of groups called clusters. Each observation belongs to a single cluster C_k. The goal of this technique is to create clusters where variations between observations are as small as possible.

Methodology

If we write $V(C_k)$ the measure of the variation within cluster k, the goal of K-Means clustering is:

$$\min_{C_1,\dots,C_K} \sum_{k=1}^{K} V(C_k) \tag{1}$$

so that the sum of within-clusters variations is the smallest possible. In order to solve this minimization problem, you first need to define a way to measure within-clusters variations. The most common but also the most suitable for FC analysis involves *squared Euclidian distance*. In this case, within-cluster variation is defined as follows:

$$V(C_k) = \frac{1}{n(C_k)} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2, \tag{2}$$

where $n(C_k)$ is the number of observations in the kth cluster and x_{ij} is the performance of city i on criterion j. This equation can be rewritten as:

$$V(C_k) = 2 \sum_{i \in C_k} \sum_{j=1}^{p} (x_{ij} - \bar{x}_{kj})^2, \tag{3}$$

with $\bar{x}_{kj} = \sum_{i \in C_k} x_{ij}$, the mean for criterion j in cluster C_k. The vector of p criteria means $(\bar{x}_{k1}, ..., \bar{x}_{kp})$ for cities included in a cluster k is called its *centroid* and will play a crucial role in the optimization algorithm. Combining equation (1) and (3), we get the following optimization problem:

$$\min_{C_1,...,C_K} 2 \sum_{k=1}^{K} \sum_{i \in C_k} \sum_{j=1}^{p} (x_{ij} - \bar{x}_{kj})^2. \tag{4}$$

Now that we explained and developed the problem, we will turn to the algorithm that allows us to find a local optimum.

1. Choose a number K of clusters you want to form with your N financial centres.
2. Randomly assign each financial centre to one of the K clusters.
3. Iterate the following procedure until clusters are fixed:
   (a) For each cluster, compute its centroid.
   (b) Assign each financial centre to the cluster whose centroid is closest (closest being defined as Euclidian distance).

Benefits and Drawbacks

The main problem of this method is linked to the *a priori* selection of a number of clusters. We will see in the next section that the hierarchical clustering technique solve this issue. Another way to find a suitable number of clusters is to use the *elbow method*. You need to compute *K-means* algorithm for different values of k and calculate the *total within-cluster sum of squares*. Plot its value against the number of selected clusters. You should be able to see that increasing k diminish drastically the total within-cluster sums of squares at first but the effect

becomes increasingly marginal. At some point you should see an "elbow" in the graph, indicating the number of clusters that should be kept.

Another drawback is that the algorithm we presented leads to a local optimum, which means that it might not be the true optimum. To deal with this issue, James et al. (2013) advise to run the algorithm several times, starting with different initial configurations and then to select the solution with the smallest objective function. It is also possible to implement an algorithm called *K-means ++* that differs from *K-means* in the way it chooses initial clusters. Arthur and Vassilvitskii (2007) show that it is both more accurate and faster.

On the other hand, the K-Means clustering method stands out as a simple and intuitive way to partition observations into several groups. Compared to Multiple-Criteria Decision Analysis techniques, assumptions are very small and it is a good way to let the data speak for themselves.

3.2.2 Hierarchical Clustering

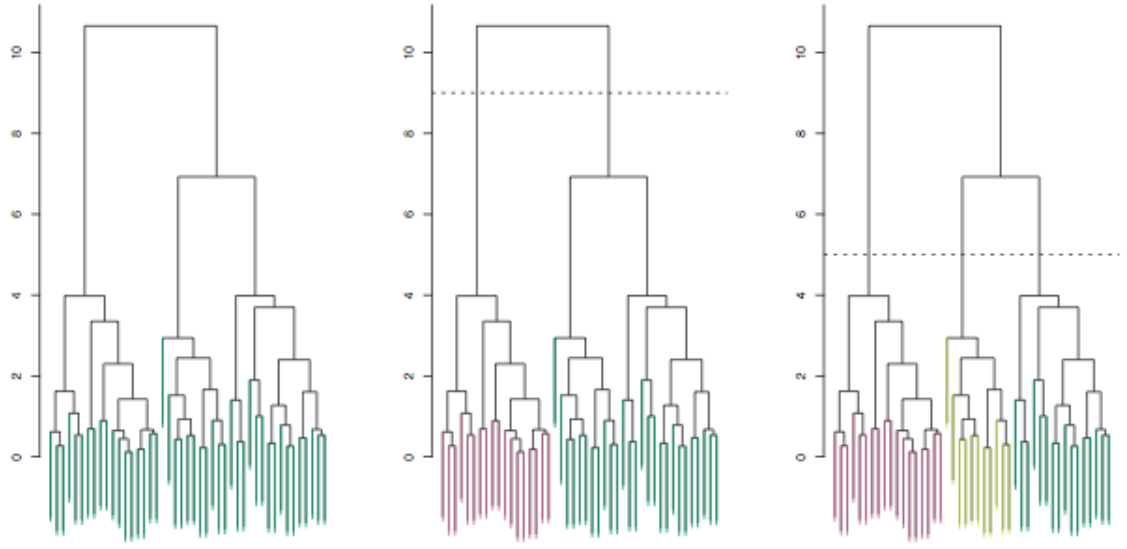
Hierarchical clustering, unlike K-Means clustering, allows to classify observations into a number of clusters that you do not need to choose before running the algorithm. Instead, it is an *a posteriori* choice that is made on the basis of a graphical representation called *dendogram*. We will see how to build it and to interpret it to create clusters.

Methodology

To understand the methodology of hierarchical clustering, it is easier to start with the end. Figure 1 shows a dendogram, which is the graphical representation of the result of this method and can be thought of as a tree. Each coloured line is called a *leaf* and represents an observation (a FC in our case). The scale on the left displays the level of dissimilarity between two components. Thus, the closer to 0 two cities (or one city and one cluster) fuse, the more similar they are. We will see how this similarity is calculated. However, please note that the proximity of two leaves on the horizontal axis has no meaning, and you should only focus on where they fuse on the vertical axis. Another major feature of the dendogram is that it is not cut in several clusters at first. As can be seen from left to right in figure 1, there is one cluster at first that you can decide to split into two or three, depending on how you interpret results. You could even go further and decide to divide the

financial system into 4,5,6 or as much clusters as there are observations, but in this case there seems to be a break after 3. Nevertheless, the choice is entirely up to the person implementing this method.

Figure 1: Same dendogram not splitted and splitted into two or three clusters, depending on the height where you decide to cut the tree.



Source: James et al. (2013)

Now we turn to the explanation of the underlying algorithm allowing to create this kind of dendogram. We first need to select a dissimilarity measure to compare pairwise clusters (where a cluster is at first composed of a single city). As with K-Means clustering, we recommend to use the Euclidian distance. *Correlation-based distance* can also be used[5] if your goal is to identify FC that behave the same way on the criteria you selected rather than to group FC according to their absolute performance. However, the concept of dissimilarity we just selected can only be applied to a comparison between two FC, but not at a higher stage where we will need to compare a FC to a group of FC already fused in a cluster. To address this problem, we need to use the concept of *linkage*, which allows to calculate the dissimilarity between two groups of FC. Literature provides four common types of linkages that are detailed in the appendix, but two of them (average and complete) seem to yield better results. Once again, it is recommended to use several linkages

[5]More informations in James et al. (2013).

in order to test the robustness of your results. Now that we laid out all possibles settings, we will present the algorithm.

1. Start with your N financial centres that will be treated as N clusters.
2. Compute the $N(N-1)/2$ pairwise dissimilarities.
3. Iterate the following steps for $i = N, N-1, ..., 2$:
   (a) Go through all pairwise dissimilarities among the i clusters and find the smallest one (*i.e.* the two most similar clusters). Fuse these two clusters and report the dissimilarity level on the vertical axis of the dendogram.
   (b) Compute again the new pairwise dissimilarities among the remaining clusters.

Benefits and Drawbacks

As we already saw, the main benefit of hierarchical clustering is that it overcomes the main drawback of K-Means clustering. Indeed, you do not need to specify a targeted number of clusters. Besides, as it results in a tree-like structure, it allows to decide *a posteri* the number of clusters. Furthermore, the selection of a dissimilarity measure as well as a type of linkage can seem a bit tricky at first glance, but enable to get different insights on the data you are analyzing.

One of the clear drawbacks of this method is that it does not give a unique classification of clusters. The different specifications of the algorithm, as it may give interesting insights, may also confuse the builder in search of a "true" answer. Of course, despite its name, this method is also not suitable to get a clear hierarchisation of clusters. It may be possible to infer what cluster is "superior" to the other, depending on the data you used, but it is not the goal of hierarchical clustering.

3.3 Factor Analysis

Factor analysis is a statistical method that allows to map a small number of unobserved factors to variables studied. As it is a way to simplify and characterize a set of data, it is used in many fields. For the purpose of this paper, we will focus on one application of this method that is quite marginal: the use of factor scores to rank observations. The easiest way to perform factor analysis is to use SPSS software from IBM but it is also possible to implement it with other statistical programs, such as the *psych* package in R[6]. In this section, we will explain the basic methodology followed by SPSS, but also provide enough references to depart from it in many ways. First and foremost, we should be more formal and define some terms.

As we just said, in factor analysis, each observed variable i is a linear sum of j unknown factors and an error term (as opposed to Principal Component Analysis where components are a linear sum of variables):

$$Y_i = \beta_{i1}F_1 + \beta_{i2}F_2 + ... + \beta_{iJ}F_J + e_i, \tag{1}$$

where β_{i1} is called the *loading* of variable i on factor 1. From now on, our goal is to construct the variance-covariance matrix and we must assume two things. Firstly, the error terms are independent of one another, and such that $\mathbb{E}(e_i) = 0$ and $Var(e_i) = \sigma^2$. Secondly, factors are independent of one another and of the error terms, and such that $\mathbb{E}(F_j) = 0$ and $Var(F_j) = 1$. Thus, we can write the variance of each observed variable as:

$$\begin{aligned} Var(Y_i) &= \beta_{i1}^2 Var(F_1) + \beta_{i2}^2 Var(F_2) + ... + \beta_{iJ}^2 Var(F_J) + Var(e_i) \\ &= \beta_{i1}^2 + ... + \beta_{iJ}^2 + \sigma^2. \end{aligned} \tag{2}$$

Here we can outline an important feature of factor analysis. Indeed, $\beta_{i1}^2 + ... + \beta_{iJ}^2$ is called the *communality*, which is the part of the variance explained by the common factors. We will try to maximize the communality and, on the other hand, to minimize the *specific variance* σ^2. We can express the covariance between

[6]Please refer to the linked chapter for a detailed explanation.

variable Y_i and variable Y_k as:

$$\begin{aligned} Cov(Y_i, Y_k) &= \beta_{i1}\beta_{k1}Var(F_1) + \beta_{i2}\beta_{k2}Var(F_2) + ... + \beta_{iJ}\beta_{kJ}Var(F_j) \\ &= \beta_{i1}\beta_{k1} + \beta_{i2}\beta_{k2} + ... + \beta_{iJ}\beta_{kJ}. \end{aligned} \tag{3}$$

We are now able to construct the covariance matrix. As shown by Tryfos (1998), the covariance matrix of standardized variables with 0 means and standard deviations equal to 1 is the same as the correlation matrix of original variables. Thus, you can either use original variables or transformed ones. In the former case, you will simply need to compute the correlation matrix[7]. Now that we saw how it is possible to express the problem in mathematical terms, we will turn to the methodology.

Methodology

Once we constructed the theoretical covariance matrix, we will try to fit the factors loadings to observed data. In our case, data are variables characterizing financial centres. From these variables, we can construct the observed variance-covariance matrix. The most commonly used technique[8] to fit theory to the observed data is the *principal components method*, which aims at finding loadings values that bring the total communality as close to the total observed variances as possible. For a detailed explanation of this method as well as other ones (such as the *principal factor*, *iterated principal factor* and *maximum likelihood* methods), you can refer to Rencher and Christensen (2012). There are also several ways to determine the number of factors J to use. By default, SPSS use the *MinEigen* method, which consists in extracting a number of factors equals to the number of eigenvalues greater than the average eigenvalue. If you used the correlation matrix, it means J will be equal to the number of eigenvalues greater than 1. A common way to confirm this choice is to make sure that the variance accounted for by this J-factors model is superior to a predetermined percentage (usually 80%-90%) of the total variance.

The next choice that needs to be made is about the *rotation* procedure to apply to the initial loadings. Indeed, there exists an infinity of possibilities for the initial factors values that fit the observations equally[9]. Thus, you could prefer

[7] This is precisely what SPSS does.
[8] By default in IBM's SPSS software.
[9] You can find the proof in Rencher and Christensen (2012) or Tryfos (1998).

another set of factor loadings (usually for interpretation reasons) to the one your first computed. There are two kinds of rotations: *orthogonal* and *oblique*. Using an orthogonal rotation, we keep factors strictly independent of one another, whereas oblique rotation allows for correlation between them. The most commonly used orthogonal rotation method is the *varimax criterion*, which finds factors that load highly on some variables and not on others. It's a good way to interpret factors if you already have an idea of what they could represent and believe they are not correlated. However, as pointed out by Osborne (2015), there is no risk to resort to oblique rotation, as it does not force factors to be correlated[10]. It can also be a good way to make sure that factors are independent, if you first opted for an orthogonal rotation.

To be able to rank cities, we need to compute another matrix: the *Component Score Coefficient Matrix*. We will describe the *regression* method, implemented by default in SPSS software. Its aim is to adjust factor loadings to take into account the initial correlations between observed variables. To get it, we simply multiply the rotated factor loadings matrix by the inverse of the initial correlation matrix. Now that we have the component score coefficient matrix, we can calculate the score of each city on each factor.

If FL_{ij} is the loading of variable i on factor j and $u_a(Y_i)$ the performance of city a on variable i, we can compute the score of city a on factor j, before aggregating factors to get the global score:

$$S_a(F_j) = \sum_{i=1}^{I} FL_{ij} u_a(Y_i) \tag{4}$$

$$S_a = \sum_{j=1}^{J} S_a(F_j). \tag{5}$$

The breakdown of the global score by factor score for each city allows us to create a ranking per factor. Thus, provided that we can interpret each factor, we can see on which dimension of the ranking one city performs better than the others. To conclude, we can summarize the methodology we just presented with the following enumeration.

1. Compute the covariance/correlation matrix, wether or not you first standardized data.

[10]SPSS offers two oblique rotation methods (*direct oblimin* and *Promax*).

2. Determine the number of factors to use and compute the first set of factor loadings, using for instance the principal components method.

3. Rotate the factor loadings according to the best suited method.

4. Compute the component score coefficient matrix.

5. Calculate each city score on each factor. Sum factors to get the global score.

6. Rank clusters in decreasing order according to the value of the score obtained.

Benefits and Drawbacks

The factor analysis method applied to clusters ranking has several benefits. First, it leads to global scores that allow us to hierarchically rank cities, but it also provides factor scores that can be useful for the analysis of strengths and weaknesses of each cluster. It is also a quite tractable method, as many settings are customizable. This tractability lets the analyst adapt the model according to his own interpretation of data. Finally, it is a methodology that has proved its worth, particularly in the social sciences, but has not been implemented yet for this specific purpose. Hence, it could yield interesting results to be compared to existing methodologies.

The main drawback of this method is related to the interpretation of factors. Indeed, there is no guarantee that factor loadings will be easily understandable and even though the goal of the rotation step is to give more sense to these loadings, it is sometimes not straightforward. It is also essential to know perfectly your dataset in order to choose the right settings in the first three steps. Lastly, if you do not have access to IBM's SPSS software, it will require more efforts to implement factor analysis on another statistical program, although R offers comprehensive tools as mentioned in the introduction of this part.

3.4 Network Analysis

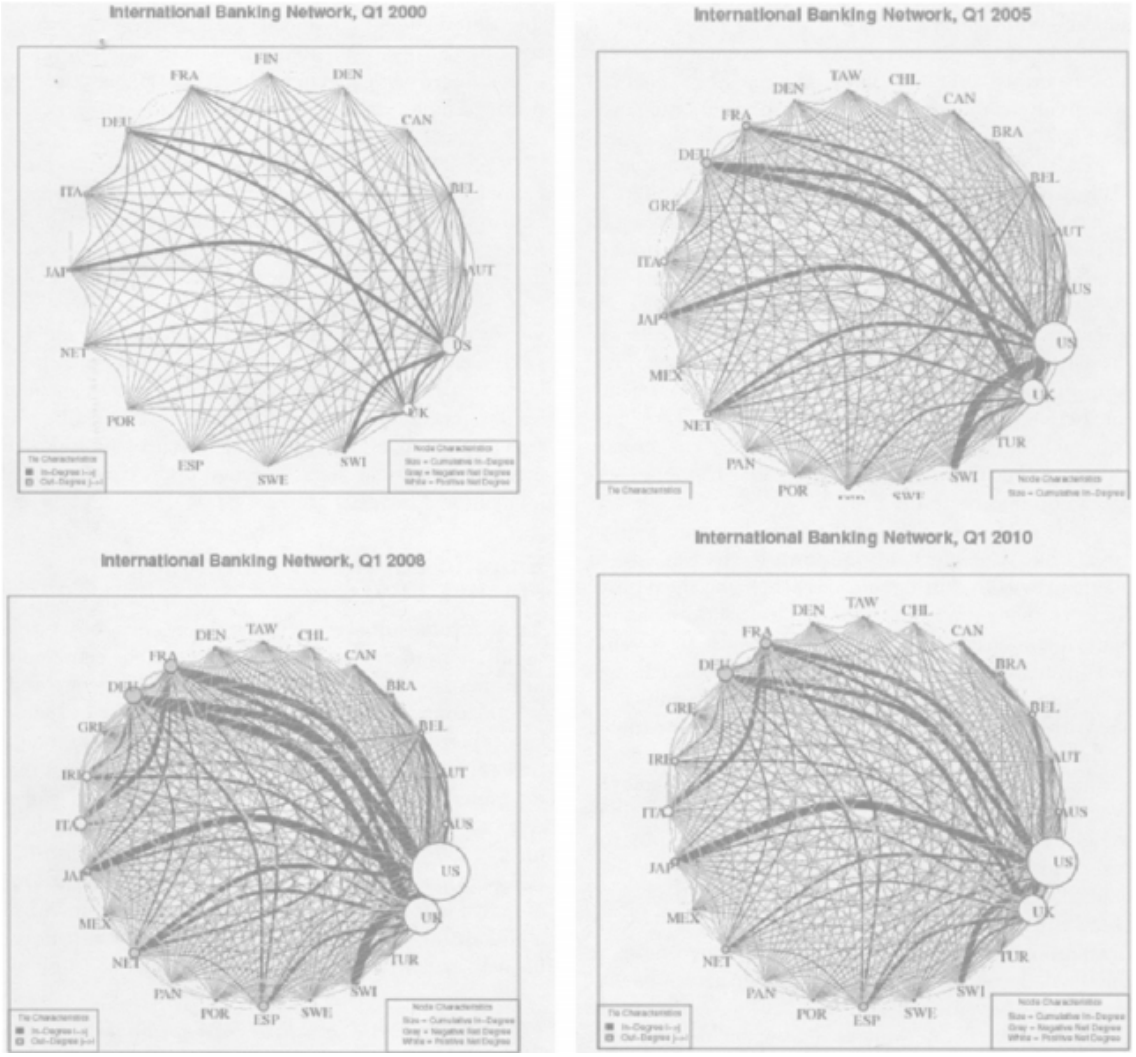
Instead of seeing FC as independent places competing with one another, we can see the financial system as an interconnected network of clusters (*nodes* in network theory vocabulary) that are all tied to each other. To do so, we can draw on the insights provided by the field of economic geography. As Tordoir (2013) showed, the international financial system match the definition of network proposed by Castells (2009). Indeed, it is an interconnection of nodes that interact with each others on different levels. They are tied by the subsidiaries of financial actors but also by trade, flows, social or politics interactions. The more information and flows the city processes, the bigger the node. Nodes can be characterized at different levels: cities for our research question but also at lower (corporates) or higher (countries) levels. According to Oatley et al. (2013), the international financial system can be theoretically described as a *complex network*, which means that its "structure is irregular, complex and dynamically evolving in time".

Methodology

In this section, we will briefly present the methodology used in Oatley et al. (2013) and give some ideas to adapt it to FC classification. In their paper, they use macroeconomic data because they are interested in the global financial system, with nodes at a national level. To construct the network presented in figure 2, they gathered data on the quantity of bank assets held by each country in foreign countries. Then, they drew lines from one country to another to represent asset holdings. Assuming that country i comes before country j in alphabetical order, if i holds bank assets in j, a black line (called *tie*) connects the two nodes. If j holds bank assets in i, a grey line is drawn. A white node corresponds to a country where inflows from abroad are superior to outflows (and vice versa for a grey node). The size of nodes and the thickness of ties are proportional to the size of holdings and allow the authors to rank countries in terms of importance in the global financial system, even though it is not their primary goal. Drawing the graph at different dates is also a good opportunity to see how nodes and relations between them evolved over time. The authors implemented this methodology thanks to packages *tnet* and *igraph* in R[11].

[11] See this link and Csardi and Nepusz (2006).

Figure 2: International banking network at different dates.



Source: Oatley et al. (2013)

Compared to the previous methods exposed, the specificity of this one is that it only uses one criterion. Nevertheless, we can imagine drawing the same kind of graphic with multiple criteria. It would probably be less elegant but the size of the nodes would remain a key indicator of the importance of the city in the global financial network. As always, adopting a multiple criteria approach would require the normalization of variables in order to give the same weight to each criterion. You could instead adopt the graphic style of Taylor (2001), who separates nodes in order to give greater details to what is included in each of it. Studying world cities as a

network, he uses the location of three representative companies to characterize the size of a node. In the case of FC analysis, we could use several international banks with different areas of expertise. If you can find relevant data, you may even be able to draw a graphic similar to the one in Taylor (2001), but using the sum of flows to determine the size of each node as in Oatley et al. (2013). Another possibility is to draw several graphs, each one based on a specific criterion. If you retain one criterion to represent one dimension of what people expect from a financial cluster (following the analytical hierarchy process for instance), you can build a ranking for each dimension and then aggregate them to find a global score. These are only a few examples of what could be done by combining the network approach to already implemented methodologies to classify FC. It is clearly not the only way to apply this methodology to our research question and the novelty of the approach offers a wide range of possibilities.

Benefits and Drawbacks

Compared to the methodologies we presented until then, network analysis has several benefits. First, it offers an elegant graphical representation that allows to see the level of interactions between FC. The great advantage over clustering methods is that it also provides a hierarchical ranking. It is also a useful tool to analyze the evolution of FC importance over time. Indeed, plotting graphs at different dates highlights the progressive reorganization of financial activities (it could be interesting to use this kind of approach to see the effects of Brexit over the years). Finally the novelty of the approach enables researchers to envision original and insightful studies. The study of complex networks being a recent area of research, it is yet to be applied to most of financial economics issues. In the case of international FC, the literature is growing but there are still many issues to tackle with these tools.

The main drawback of this method is the difficulty to get data. To characterize ties between nodes, you will need data either from banks' location or trade flows if you want to apply this method to FC. These data are presumably more complicated to gather than macroeconomic data like the ones used in Oatley et al. (2013). A good source of information could be *The Banker*'s database compiled by the *Financial Times*.

4 Conclusion

In this article, we spanned a wide range of methodologies that can be used either to classify FC into groups or to rank them. As we already emphasized, each method has its own implications and differs in the output it produces. Some of them have been extensively implemented in the literature, while others have barely been applied to FC classification, and thus offer a lot of possibilities. They also all vary in their degree of complexity. All these specificities need to be considered before starting to build a classification.

After choosing a method consistent with the characteristics of a financial cluster you would have defined beforehand, you can start selecting data accordingly. Most of the time, these data will need to be prepared: restated to ensure the uniformity of the dataset, rescaled to enable aggregation of variables or processed to fix correlation issues. All these procedures need to be clearly stated to guarantee the transparency of the methodology. The results you will get will depend on these choices, so it is always better to test their robustness by modifying some settings and see how it affects your findings. It can also be interesting to implement several methods, either to compare results or to obtain complementary outputs (one leading to a ranking and the other to a graphical representation for instance).

Appendix

Saaty's pairwise comparison scale

Importance	*Definition*	*Explanation*
1	Equal importance	Two activities contribute equally to the objective
2	Weak importance	
3	Moderate importance	Experience and judgment slightly favor one activity over another
4	Moderate plus importance	
5	Strong importance	Experience and judgment strongly favor one activity over another
6	Strong plus importance	
7	Very strong or demonstrated importance	An activity is favored very strongly over another; its dominance demonstrated in practice
8	Very, very strong importance	
9	Extreme importance	The evidence favoring one activity over another is of the highest possible order of affirmation

Source : Saaty and Vargas (2012).

Average random consistency index

N	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Index	0.00	0.00	0.52	0.89	1.11	1.25	1.35	1.40	1.45	1.49

Source : Saaty and Vargas (2012).

Types of linkage for hierarchical clustering

Linkage	***Description***
Complete	Maximal intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the largest of these dissimilarities.
Single	Minimal intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the smallest of these dissimilarities. Single linkage can result in extended, trailing clusters in which single observations are fused one-at-a-time.
Average	Mean intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the average of these dissimilarities.
Centroid	Dissimilarity between the centroid for cluster A (a mean vector of length p) and the centroid for cluster B. Centroid linkage can result in undesirable inversions.

Source : James et al. (2013).

References

K. J. Arrow. *Social choice and individual values*. 1951. 4

D. Arthur and S. Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics, 2007. 20

J. Boissinot and T. Marx. What makes a financial center attractive? : A literature review. *Not published yet*, 2017. 3

M. Castells. *Communication power*. New York: Oxford University Press, 2009. 27

G. Csardi and T. Nepusz. The igraph software package for complex network research. *InterJournal, Complex Systems*, 1695(5):1–9, 2006. 27

G. James, D. Witten, T. Hastie, and R. Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013. 9, 18, 20, 21, 32

Karmeshu. *Entropy measures, maximum entropy principle and emerging applications*, volume 119. Springer Science & Business Media, 2003. 9

M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938. 8

S. Kullback and R. A. Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951. 9

M. Nardo, M. Saisana, A. Saltelli, S. Tarantola, A. Hoffman, and E. Giovannini. Handbook on Constructing Composite Indicators: Methodology and User Guide. OECD Statistics Working Papers 2005/3, OECD Publishing, Aug. 2005. 5

T. Oatley, S. B. Danzman, A. Pennock, and W. K. Winecoff. Political economy of global finance: A network approach. *Perspectives on Politics*, 11(1):133–153, 2013. 27, 28, 29

J. W. Osborne. What is rotating in exploratory factor analysis. *Practical Assessment, Research & Evaluation*, 20(2):1–7, 2015. 25

K. Pearson. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58:240–242, 1895. 7

A. C. Rencher and W. F. Christensen. Exploratory factor analysis. *Methods of Multivariate Analysis, Third Edition*, pages 435–477, 2012. 24

B. Roy. Classement et choix en présence de points de vue multiples. *Revue française d'informatique et de recherche opérationnelle*, 2(8):57–75, 1968. 16

B. Roy. Electre 3 : Un algorithme de classement fondé sur une représentation floue des préférences en présence de critères multiples. *Cahiers du CERO*, 20(1):3–24, 1978. 16

B. Roy and P. Bertier. La méthode electre 2 (une application au média-planning...). 1973. 16

T. L. Saaty. What is the analytic hierarchy process? In *Mathematical models for decision support*, pages 109–121. Springer, 1988. 13

T. L. Saaty and L. G. Vargas. *Models, methods, concepts & applications of the analytic hierarchy process*, volume 175. Springer Science & Business Media, 2012. 13, 14, 31

C. Spearman. The abilities of man. 1927. 7

P. J. Taylor. Specification of the world city network. *Geographical analysis*, 33(2):181–194, 2001. 28, 29

S. Tordoir. Network theory applied to the global financial system: a new paradigm. *Atlantic-community. com*, 2013. 27

P. Tryfos. *Methods for Business Analysis and Forecasting: Text and Cases*. Wiley, 1998. 24