

## Introduction

**Cluster analysis** or **clustering** is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters).

The idea is to maximize between group variance and, at the same time, minimize within group variance. In cluster analysis, there is no prior information about the group or cluster membership for any of the objects.

Clustering is mostly applied to complex phenomenon. In this case complex phenomenon is understood as multidimensional, and in many cases abstract concept, which cannot be captured by single indicators (for example: quality of life, sustainable development, countries competitiveness, consumer segmentation). Such a complex phenomenon can be analysed and visualized using:

- composite indicator,
- clustering method.

## Composite indicators

A composite indicator is created when individual indicators are combined into a single measure, so it may include several dimensions, where the dimensions represent different domains or aspects of the phenomenon being measured. True to its nature, a composite indicator is usually built to 'tell a story'. It is thus ideally suited to identify and bring attention to a possibly latent phenomenon.

10 steps to create a composite indicator:

- **Theoretical framework** - select variables that, after aggregation, will allow the composite indicator to achieve its purpose
- **Data selection** - data collection and data quality assessment
- Imputation of missing data
- **Multivariate analysis** - assessing the statistical and conceptual coherence in the structure of the dataset (e.g. by principal component analysis, correlation analysis, and Cronbach's alpha)
- **Normalization** - bringing indicators onto a common scale (applying one of the common normalization methods, e.g. min-max, z-scores, or the distance to best performer)
- **Weighting** - assigning individual weights that will reflect the importance of each indicator to be adjusted according to the concept being measured. Weighting methods can be statistical, based on public/expert opinion, or both.
- **Aggregation** - combining the values of a set of indicators into a single summary 'composite' or 'aggregate' measure (in the most common approach just simply by the average value of normalized indicators).
- **Sensitivity analysis** - quantifies the uncertainty caused by each individual assumption, which identifies particularly sensitive assumptions which might merit closer consideration. For example by using tools such as Monte Carlo to investigate the effects on the scores and ranks of perturbing these assumptions (i.e. alternative weighting schemes, aggregation methods, etc.)
- **Link to other measures** - identifying linkages through regressions with other existing measures
- **Visualization** - helps to effectively communicate the message

## Visualization of composite indicators

Presenting a composite indicator is not a trivial issue. Composite indicators must be able to communicate a picture to decision-makers and other end-users quickly and accurately. Visual models of composite indicators can provide signals, e.g., of problematic areas that require policy intervention.

Visualization techniques can vary from simple tabular format, line or bar chart to complicated, interactive dashboards.

Example: [OECD Better Life Index](#)

## Clustering

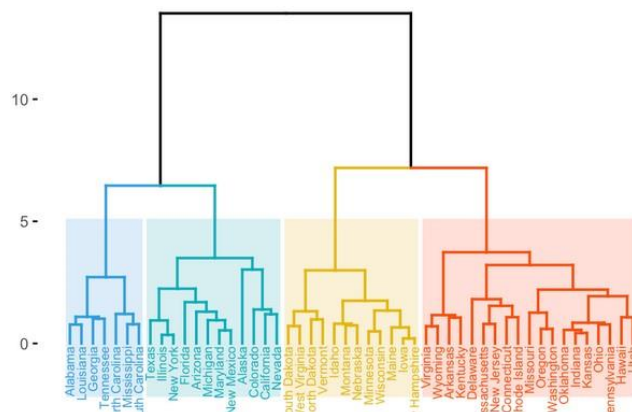
Data clustering is the process of identifying natural groupings or clusters within multidimensional data based on some similarity measure. The most popular way to evaluate a similarity measure is the use of distance measures. The most widely used distance measure is the Euclidean distance (but one can also apply Minkowski metrics, Manhattan or Mahalanobis distance).

Most clustering algorithms are based on two popular techniques known as hierarchical and partitional clustering.

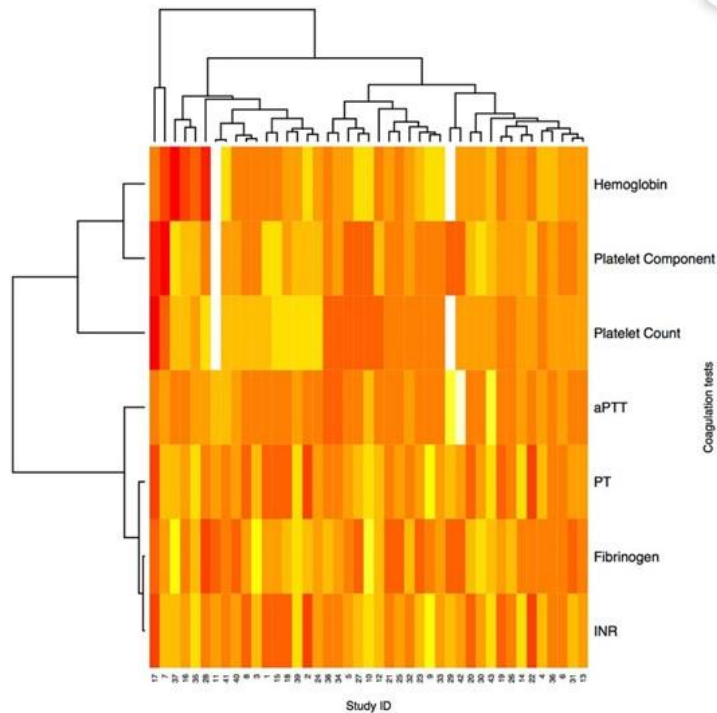
## Hierarchical clustering

Hierarchical Clustering Algorithm generate a cluster tree (or dendrogram) by using heuristic splitting or merging techniques. Divisive hierarchical algorithms start with all the patterns assigned to a single cluster. Then, splitting is applied to a cluster in each stage until each cluster consists of one pattern. Contrary to divisive hierarchical algorithms, agglomerative hierarchical algorithms start with each pattern assigned to one cluster. Then, the two most similar clusters are merged together. This step is repeated until all the patterns are assigned to a single cluster.

Visualization example:



Heatmap & dendrogram:

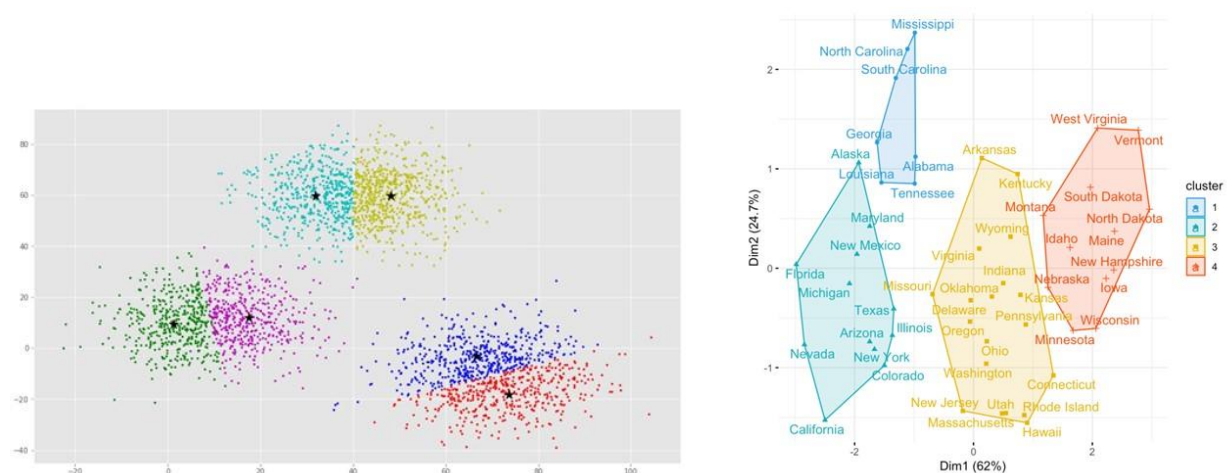


## Partitional Clustering

Partitional clustering algorithms divide the data set into a specified number of clusters. These algorithms try to minimize certain criteria (e.g. a square error function) and can therefore be treated as optimization problems. However, these optimization problems are generally NP-hard and combinatorial.

The most frequently used algorithm for clustering data is k-means algorithm. K-means starts by choosing  $k$  random centers which you can set yourself. Then, all data points are assigned to the closest center based on (in most cases) their Euclidean distance. Next, new centers are calculated and the data points are updated. This process continuous until clusters do not change between iterations.

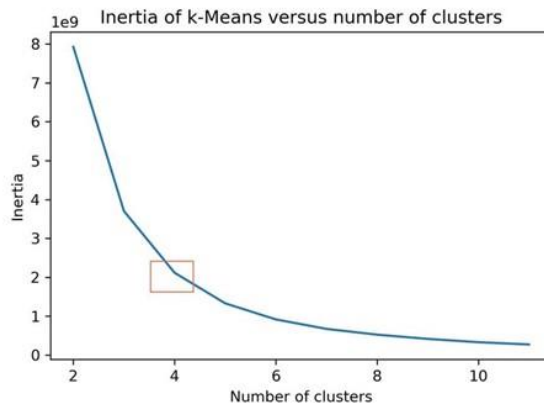
Sample cluster plots:



## K-means optimal number of clusters

The number of clusters can be set up:

1. looking first at the dendrogram,
2. applying "elbow" method



3. based on Silhouette index (recommended)

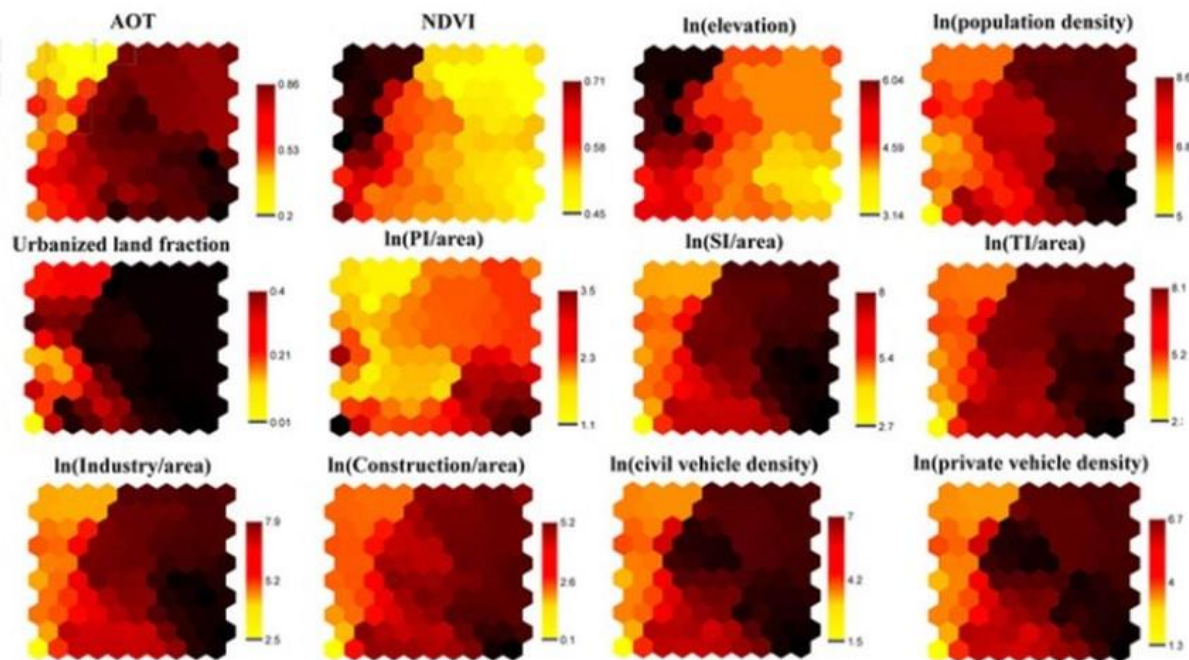
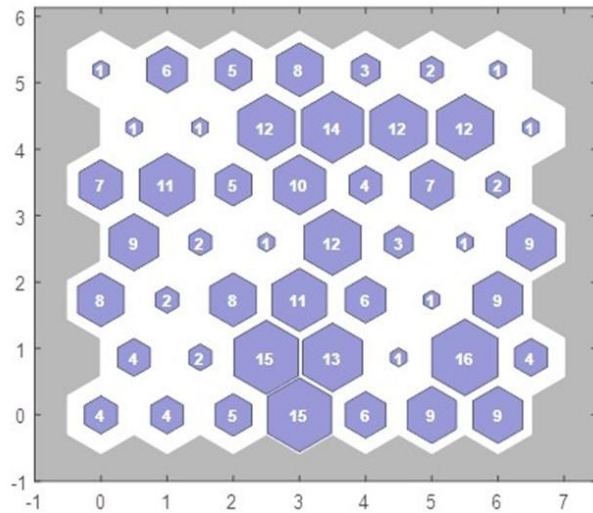
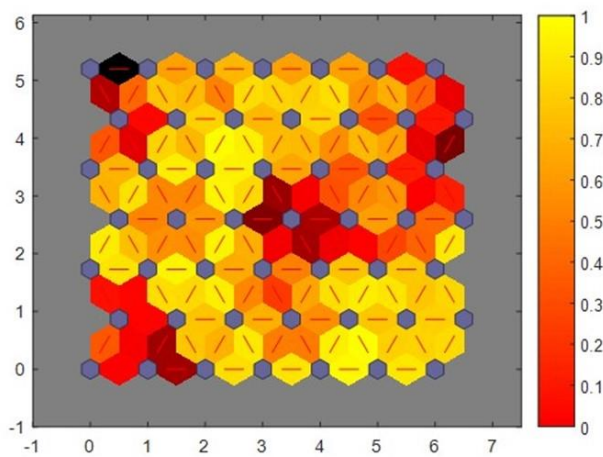
The Silhouette score measures the separability between clusters based on the distances between and within clusters.

## Kohonen's SOM

Kohonen SOMs are exploratory data analysis technique projecting multi-dimensional data onto a two-dimensional space. That procedure allows clear visualization of the data and easy identification of groups with similar characteristics. In this sense, Kohonen maps can be thought of as a factor analysis combined with a cluster analysis. The advantage of self-organizing maps over composite indicators in the classification procedure is revealed by excluding such problems as normative nature of weight, sensitivity to used normalization and aggregation method, and also a subjective selection of indicators. Another advantage of Kohonen maps is the self-organizing property of the map which makes estimated components varies in a monotonic way across the map.

Effective data clustering using SOM involves two or three steps procedure. After proper network training, units can be clustered generating regions of neurons which are related to data clusters. The basic assumption relies on the data density approximation by the neurons through unsupervised learning.

The visual outcomes of applying SOM's are **U-matrix** and **neuron location topology** (the numbers in the cells correspond to the number of elements in the clusters)



## Introduction

Symbolic data analysis has been introduced in order to solve the problem of the analysis of data given on an aggregated form, different from a central point. While classical data values are single points in  $p$ -dimensional space; symbolic data values are hypercubes (broadly defined) in  $p$ -dimensional space (and/or a Cartesian product of  $p$  distributions).




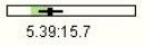





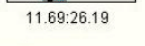

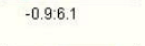

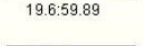

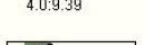

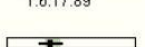

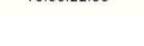
Many symbolic datasets result from the aggregation of large or extremely large classical datasets into smaller more manageably sized datasets, with the aggregation criteria typically grounded on basic scientific questions of interest. Unlike classical data, symbolic data have internal variation and structure which must be taken into account when analyzing the datasets.

Example:

u	Age	Blood Pressure	City	Type of Cancer	Gender
1	[20, 30)	(79, 120)	Boston	{Brain tumor}	{Male}
2	[50, 60)	(90, 130)	Boston	{Lung, Liver}	{Male}
3	[45, 55)	(80, 130)	Chicago	{Prostate}	{Male}
4	[47, 47)	(86, 121)	El Paso	{Breast $p$ , Lung $(1 - p)$ }	{Female}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

In fact, each cell of the data can contain:

- a number,
- a category,
- an interval,
- a sequence of categorical values,
- a sequence of weighted values,
- a bar chart,
- a histogram,
- a distribution.

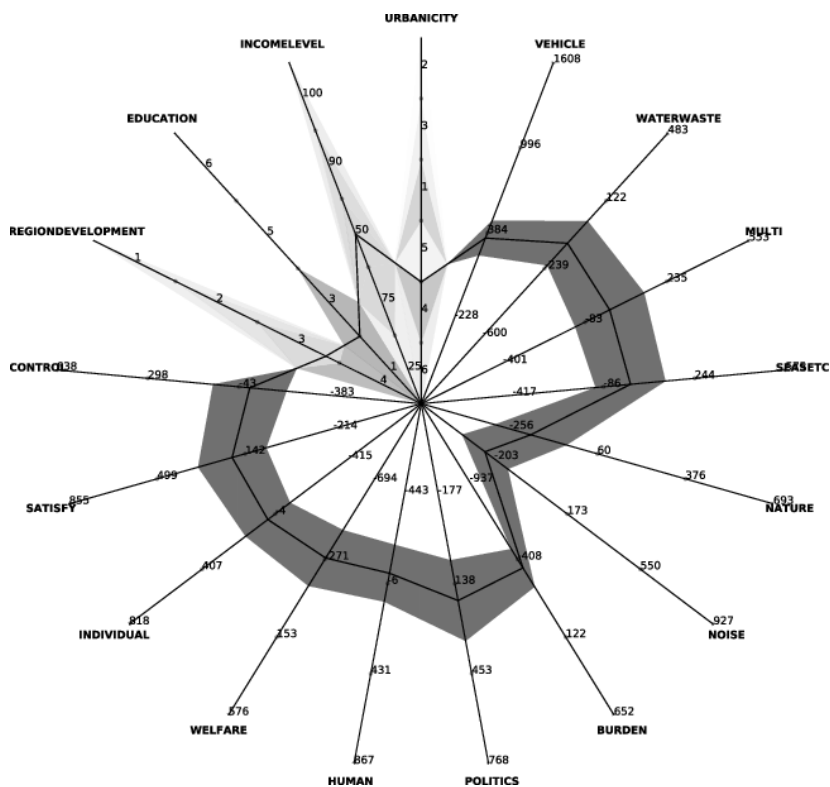
Tower	Differences_y1_y2	Settlement_Total	Sum_Cracks_y1_y2
ouv16		 13.89:23.6	5453.26
ouv17		 5.39:15.7	5300.95
ouv14		 8.3:65.7	2844.63
ouv07		 6.29:30.0	2472.83
ouv19		 11.69:26.19	4213.33
ouv09		 -0.9:6.1	4503.127
ouv08		 19.6:59.89	3268.257
ouv13		 4.0:9.39	1548.971
ouv15		 1.6:17.89	1224.28
ouv21		 13.39:22.59	-276.629

Symbolic Data are complex data as they cannot be reduced to standard data without losing much information.

## 2D Zoom Star

Zoom Star is used to represent a symbolic object. It is a radial graph where each axis represents a variable. It can be considered as an iconic technique, because the shape of the representation identifies the object. It can also be considered as a geometrical technique, like a Kiviat graph or parallel coordinates because, when the variables are quantitative, geometrical properties can be used to derive information.

The difference with glyphs, Kiviat graphs or parallel coordinates is that, here, we represent on the axis interval if the variable is quantitative or categories and frequencies if the variable is categorical.

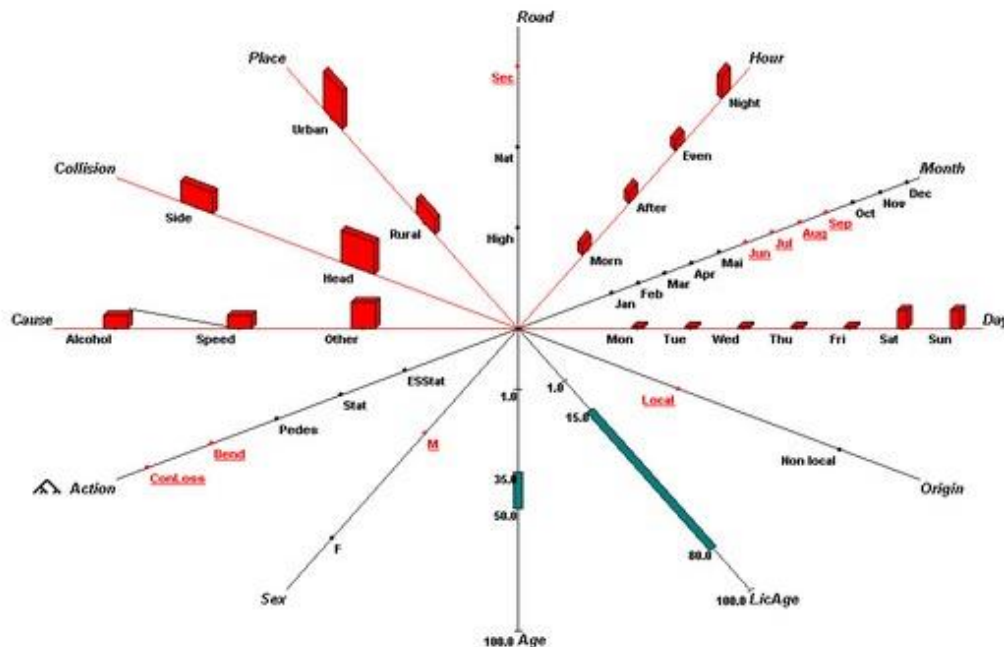


The goal of the zoom star is to provide different levels of detail in order to allow a stepwise improvement of the users' knowledge on the object. It provides a first general image of the symbolic object.



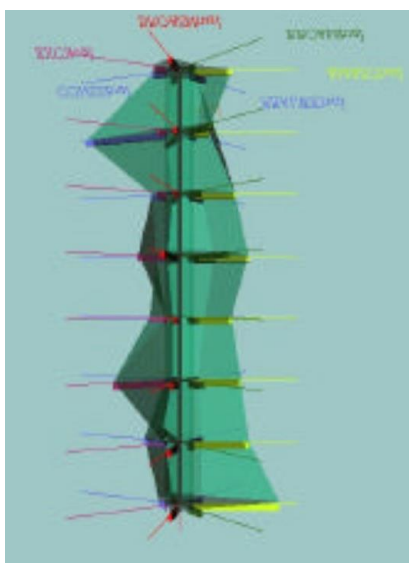
## 3D Zoom Star

In this version, the plane of the star is seen in 3D and the histograms associated with categorical variables are represented directly on the axes. It is a more detailed version than 2D Zoom Star. However, it requires a minimum of experience in order to assimilate all the information.



## Temporal star

This representation has been used to visualize a symbolic object varying with time. The 3D Zoom Stars visualizing an object at different epochs are thread on a central axis representing time. The graph can be zoomed, moved, rotated around the time axis. To emphasize the evolution from one epoch to another, the external (or internal) extremities of intervals can be joined, making an external (or internal) transparent veil.





## Pyramid

The pyramidal model generalizes hierarchies by allowing non-disjoint classes at each level, but it imposes the existence of a linear order on  $E$  such that all classes are intervals of this order.

Hence, a pyramid provides both a clustering and a seriation on the data. The pyramidal model, leading to a system of clusters which is richer than that produced by hierarchical clustering, allows for the identification of clusters that the hierarchical model would not identify; the existence of a compatible order on the objects leads, however, to a structure which is much simpler than lattices.

