HỆ HỖ TRỢ QUYẾT ĐỊNH

Bài 9 (b): Clustering

Contents

What is clustering analysis

Clustering

- Clustering is the process of finding meaningful groups in data
- For example, the customers of a company can be grouped based on purchase behavior.
- The task of clustering can be used in two different classes of applications:
  - To describe a given dataset and
  - as a preprocessing step for other data science algorithms.

Clustering to describe the data

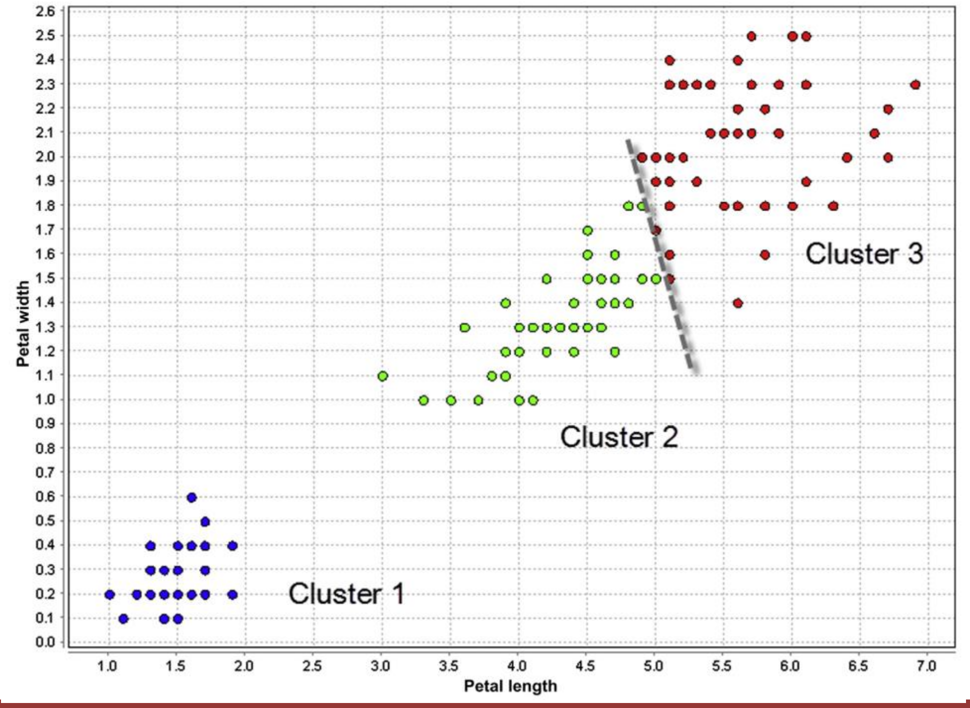
- The most common application of clustering is to explore the data and find all the possible meaningful groups in the data
- Clustering a company's customer records can yield a few groups in such a way that customers within a group are more like each other than customers belonging to a different group
- Applications:
  - Marketing: Finding the common groups of customers
  - Document clustering: One common text mining task is to automatically group documents
  - Session grouping: In web analytics, clustering is helpful to understand common groups of clickstream patterns

Clustering for preprocessing

- Clustering to reduce dimensionality
- Clustering for object reduction

Types of clustering techniques

- The clustering process seeks to find groupings in data, in such a way that data points within a cluster are more similar to each other than to data points in the other clusters
- One common way of measuring similarity is the Euclidean distance measurement in $n - dimensional$ space



Taxonomy based on data point’s membership

- **Exclusive or strict partitioning clusters**: Each data object belongs to one exclusive cluster
- **Overlapping clusters**: The cluster groups are not exclusive, and each data object may belong to more than one cluster.
- **Hierarchical clusters**: Each child cluster can be merged to form a parent cluster.
- **Fuzzy or probabilistic clusters**: Each data point belongs to all cluster groups with varying degrees of membership from 0 to 1.

Taxonomy by algorithmic approach

- **Prototype-based clustering**: In the prototype-based clustering, each cluster is represented by a central data object, also called a prototype (centroid clustering or center-based clustering)
- **Density clustering**:
  - Each dense area can be assigned a cluster and the low-density area can be discarded as noise
- **Hierarchical clustering**:
  - Hierarchical clustering is a process where a cluster hierarchy is created based on the distance between data points.
  - The output of a hierarchical clustering is a dendrogram: a tree diagram that shows different clusters at any point of precision which is specified by the user
- **Model-based clustering**:
  - Model-based clustering gets its foundation from statistics and probability distribution models; this technique is also called distribution-based clustering.
  - *Mixture of Gaussians* is one of the model-based clustering techniques

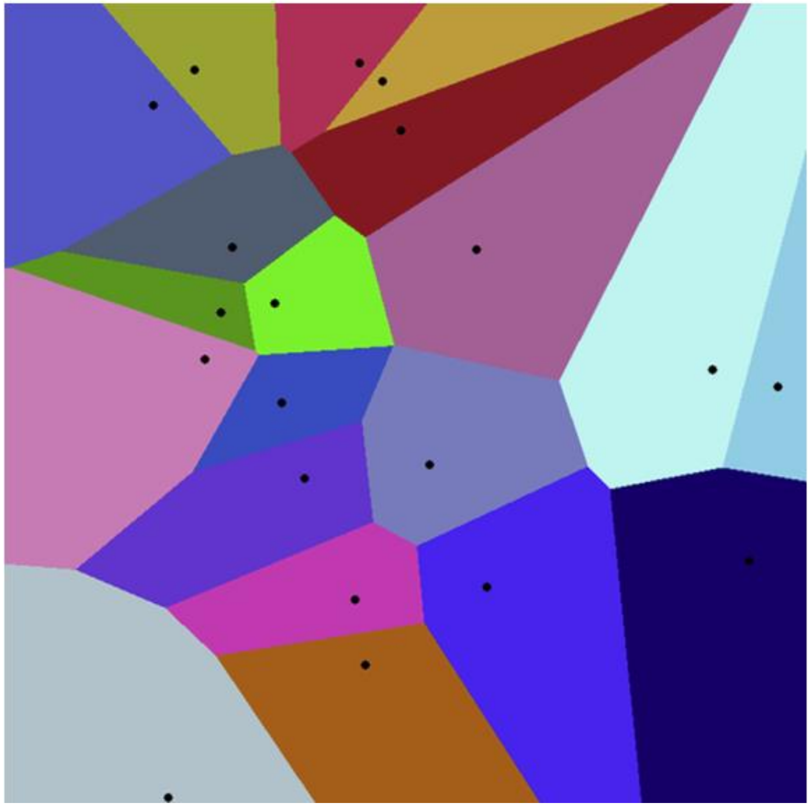
K-Means Clustering

k-Means

- k-Means clustering is a prototype-based clustering method where the dataset is divided into $k - clusters$.
- User specifies the number of clusters (k) that need to be grouped in the dataset.
- The objective of k-means clustering is to find a prototype data point for each cluster; all the data points are then assigned to the nearest prototype, which then forms a cluster.
- The prototype is called as the centroid, the center of the cluster.
- The center of the cluster can be the mean of all data objects in the cluster, as in k-means, or the most represented data object, as in k-medoid clustering.
- The cluster centroid or mean data object does not have to be a real data point in the dataset and can be an imaginary data point that represents the characteristics of all the data points within the cluster.

k-Means

- The data objects inside a partition belong to the cluster.
- These partitions are also called Voronoi partitions, and each prototype is a seed in a Voronoi partition.



Algorithm

- The logic of finding k-clusters within a given dataset is rather simple and always converges to a solution
- However, the final result in most cases will be locally optimal where the solution will not converge to the best global solution
- **Step 1**: Initiate Centroids
  - The first step in k-means algorithm is to initiate *k* random centroids
- **Step 2**: Assign Data Points
  - all the data points are now assigned to the nearest centroid to form a cluster.
- **Step 3**: Calculate New Centroids
  - New centroids are means of each clusters

Algorithm

- **<u>Step 4</u>**: Repeat Assignment and Calculate New Centroids
  - assigning data points to the nearest centroid is repeated until all the data points are reassigned to new centroids
- **<u>Step 5</u>**: Termination
  - Step 3—calculating new centroids, and step 4—assigning data points to new centroids, are repeated until no further change in assignment of data points happens.

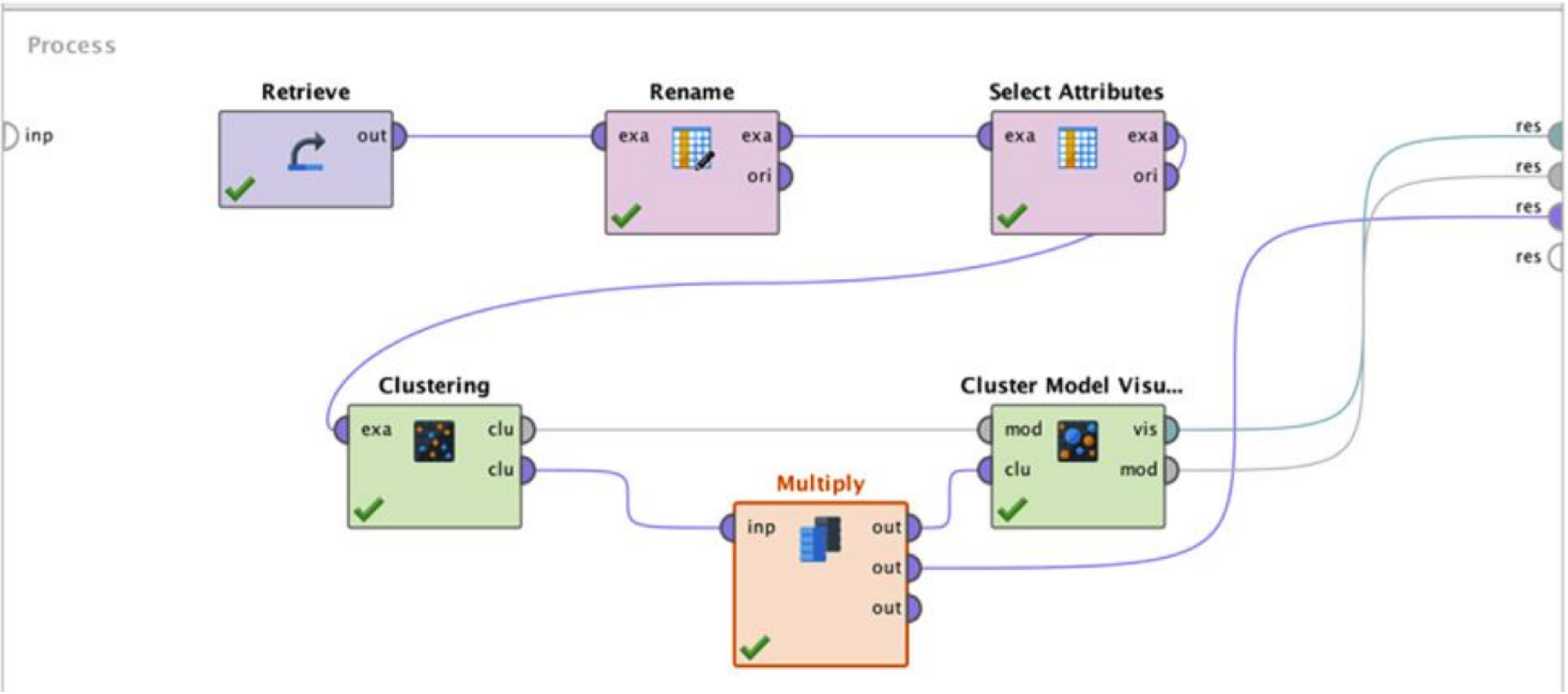
Some key issues

- Initiation: The final clustering grouping depends on the random initiator and the nature of the dataset.
- Empty clusters: One possibility in k-means clustering is the formation of empty clusters in which no data objects are associated.
- Outliers: Since SSE (sum of squared errors) is used as an objective function, k-means clustering is susceptible to outliers
- Post-processing: Since k-means clustering seeks to be locally optimal, a few post-processing techniques can be introduced to force a new solution that has less SSE

Evaluation of Clusters

- Evaluation of clustering can be as simple as computing total SSE
- Good models will have low SSE within the cluster and low overall SSE among all clusters.
- SSE can also be referred to as the average within-cluster distance and can be calculated for each cluster and then averaged for all the clusters.
- ***Davies-Bouldin*** index is a measure of uniqueness of the clusters and takes into consideration both cohesiveness of the cluster (distance between the data points and center of the cluster) and separation between the clusters

k-Means in RapidMiner



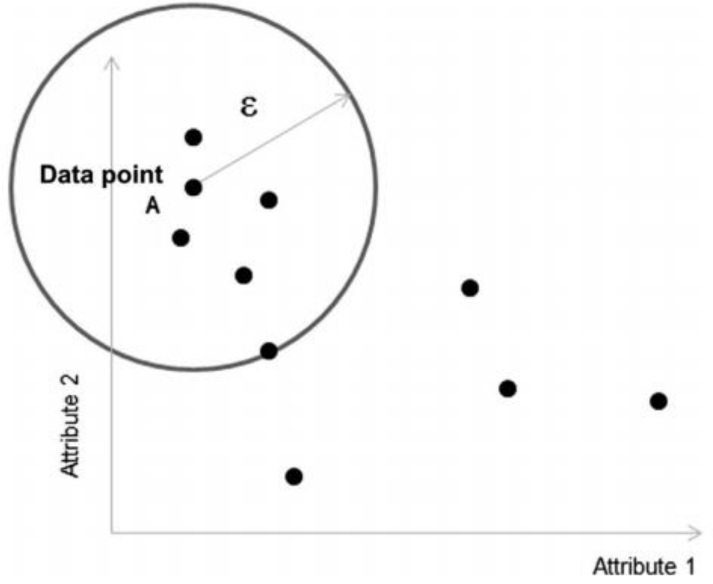
DBSCAN Clustering

DBSCAN

- A cluster can also be defined as an area of high concentration (or density) of data objects surrounded by areas of low concentration (or density) of data objects.
- A density-clustering algorithm identifies clusters in the data based on the measurement of the density distribution in n-dimensional space
- Specifying the number of the cluster parameters (k) is not necessary for density-based algorithms
- Thus, density-based clustering can serve as an important data exploration technique
- *Density can be defined as the number of data points in a unit n-dimensional space*

Algorithm

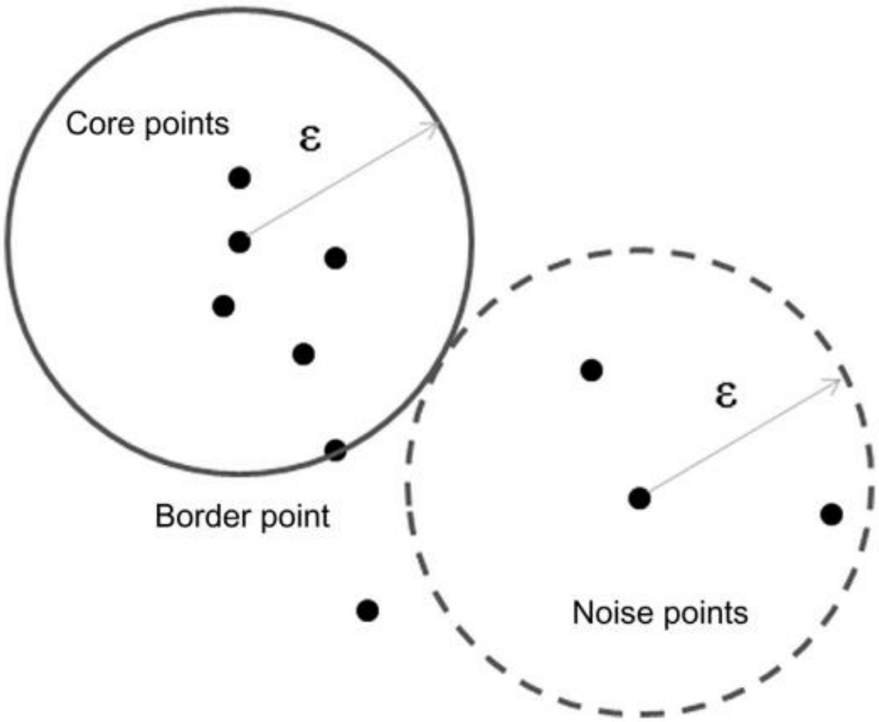
- **Step 1**: Defining Epsilon and MinPoints
  - Calculation of a density for all data points in a dataset, with a given fixed radius ε (epsilon).
  - To determine whether a neighborhood is high-density or low-density, a threshold of *data points (MinPoints)* will have to be defined, above which the neighborhood is considered high-density
  - Both ε and $MinPoints$ are user-defined parameters



Algorithm

- **<u>Step 2</u>**: Classification of Data Points

Core points: All the data points inside the high-density region of at least one data point are considered a core point. A high-density region is a space where there are at least *MinPoints* data points within a radius of ε for any data point.



Border points: Border points sit on the circumference of radius ε from a data point. A border point is the boundary between high-density and low-density space. Border points are counted within the high-density space calculation

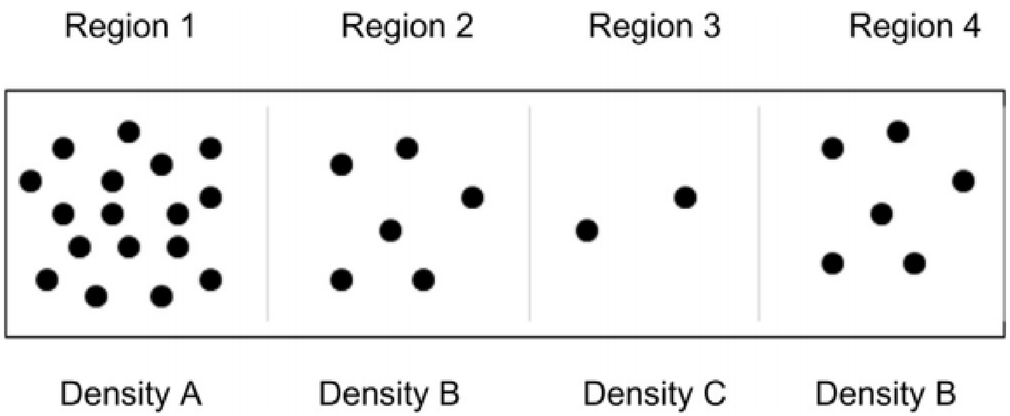
Noise points: Any point that is neither a core point nor border point is called a noise point. They form a low-density region around the high density region.

Algorithm

- **<u>Step 3</u>**: Clustering
  - Groups of core points form distinct clusters. If two core points are within ε of each other, then both core points are within the same cluster
  - All these clustered core points form a cluster, which is surrounded by low-density noise points
  - A few data points are left unlabeled or associated to a default noise cluster

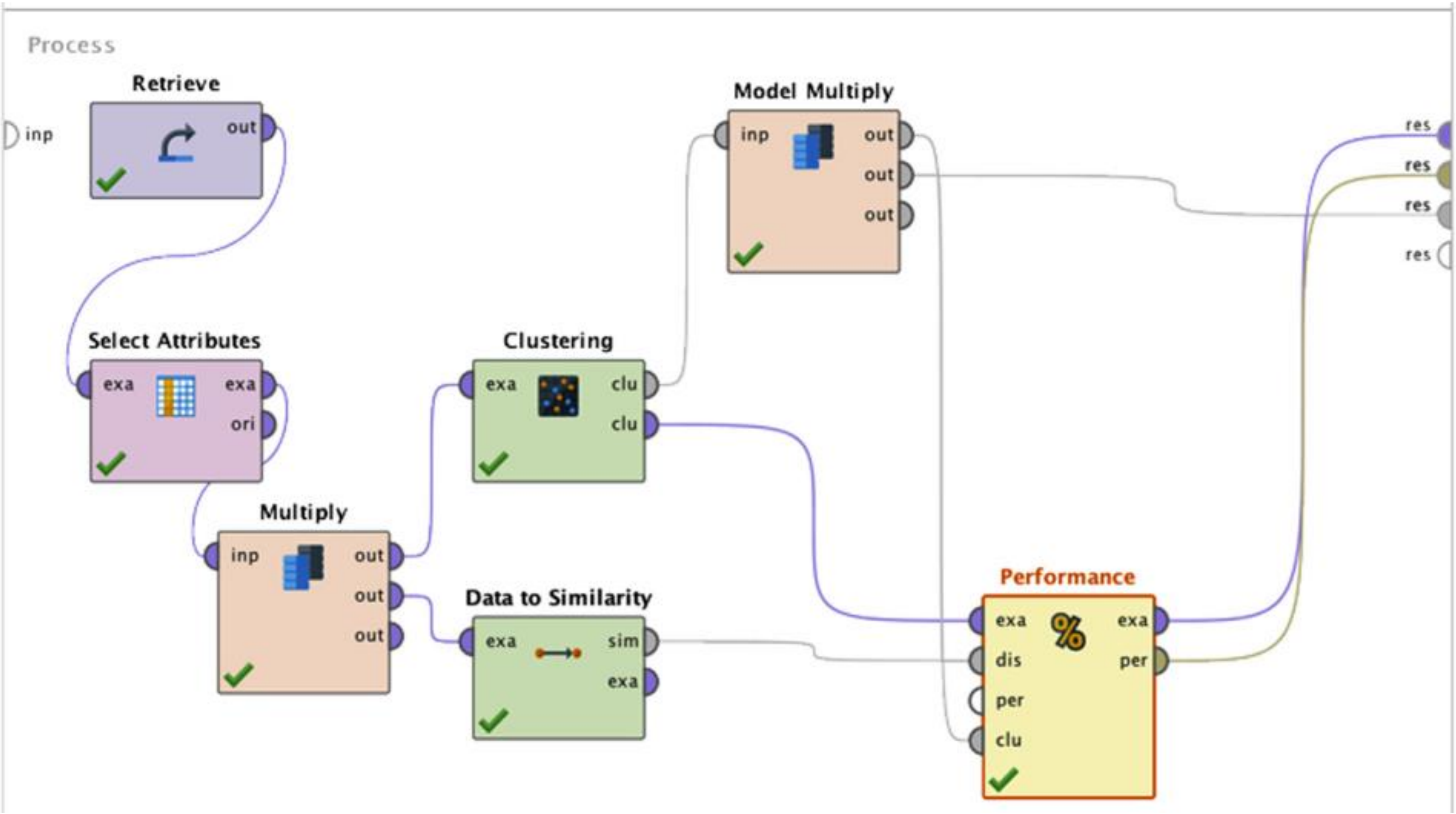
Special Cases: Varying Densities

- The dataset has four distinct regions numbered from 1-4. Region 1 is the high-density area A, regions 2 and 4 are of mediumdensity B, and between them is region 3, which is extremely low-density C



- The density threshold parameters are tuned in such a way as to partition and identify region 1, then regions 2 and 4 (with density B) will be considered noise, along with region 3

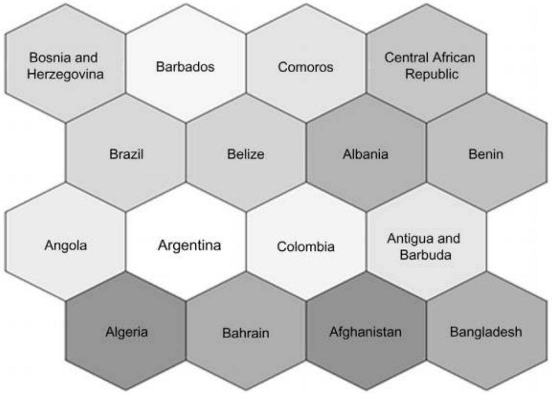
DBSCAN in RapidMiner

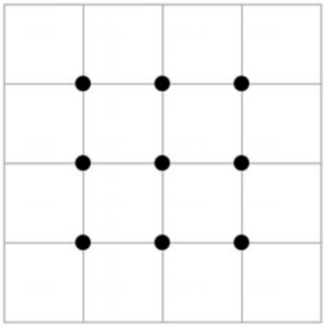


Self-Organizing Maps

SOM

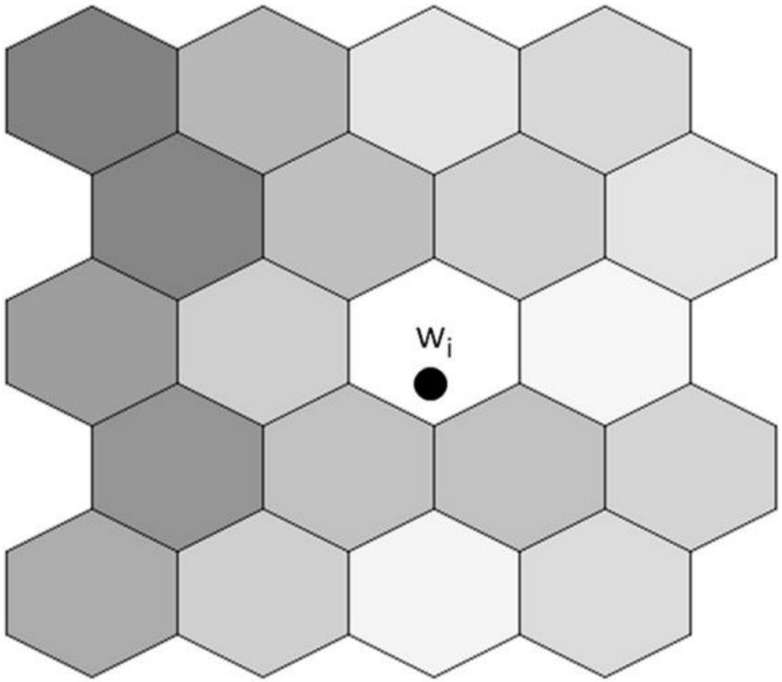
- A self-organizing map (SOM) is a powerful visual clustering technique that evolved from a combination of neural networks and prototype-based clustering
- A key distinction in this neural network is the absence of an output target function to optimize or predict, hence, it is an unsupervised learning algorithm
- SOM methodology is used to project data objects from data space, mostly in n dimensions, to grid space, usually resulting in two dimensions





Algorithm

- **<u>Step 1</u>**: Topology Specification
  - Two-dimensional rows and columns with either a rectangular lattice or a hexagonal lattice are commonly used in SOMs
  - The **number of centroids** is the product of the **number of rows** and **columns** in the grid

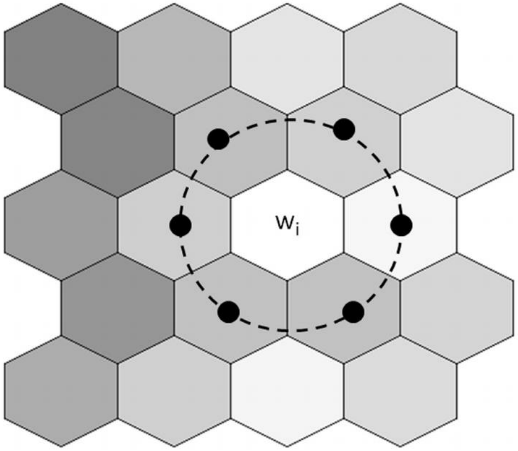


Algorithm

- **<u>Step 2</u>**: Initialize Centroids
  - A SOM starts the process by initializing the centroids. The initial centroids are values of random data objects from the dataset. This is similar to initializing centroids in k-means clustering.
- **<u>Step 3</u>**: Assignment of Data Objects
  - After centroids are selected and placed on the grid in the intersection of rows and columns, data objects are selected one by one and assigned to the nearest centroid.
- <u>Step</u> 4: Centroid Update

$$w_i(t+1) = w_i(t) + f_i(t) \times [d(t) - w_i(t)]$$

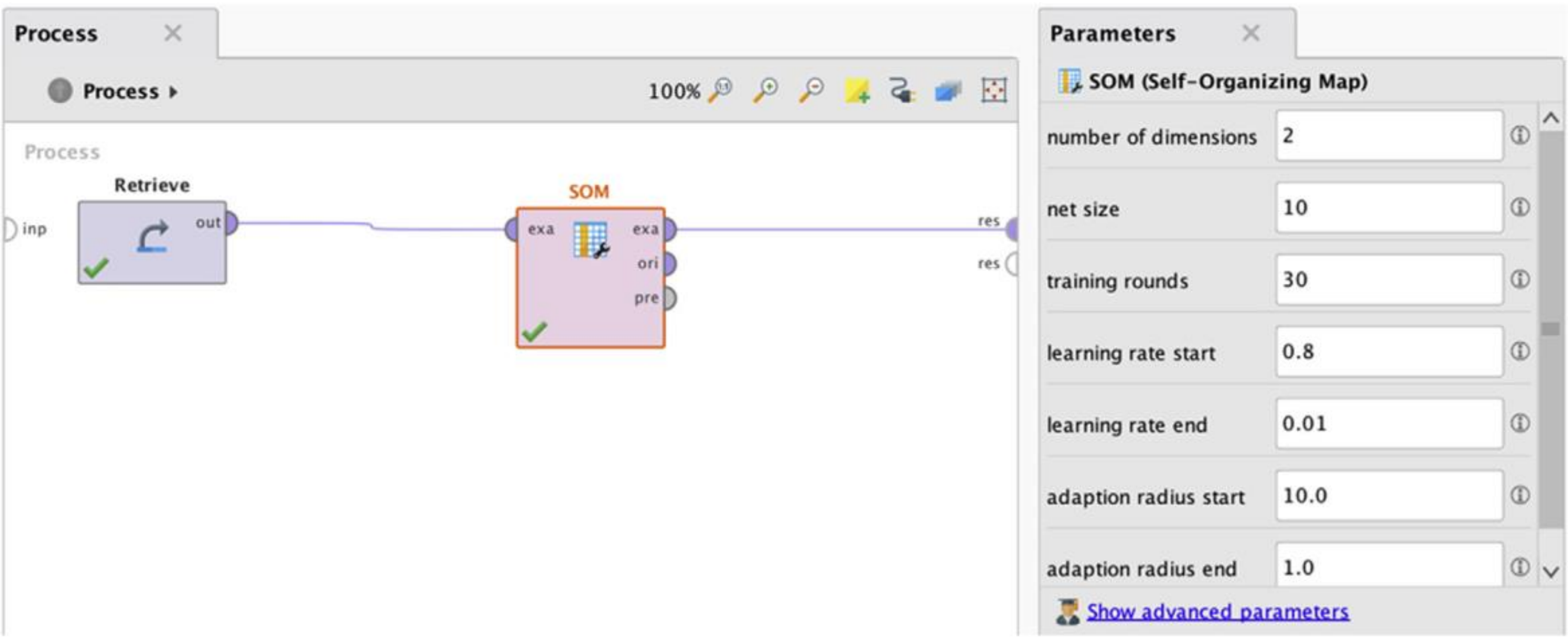
$$f_i(t) = \lambda_i(t) e^{\left(-\frac{(g_i - g_j)^2}{2\sigma^2}\right)}$$



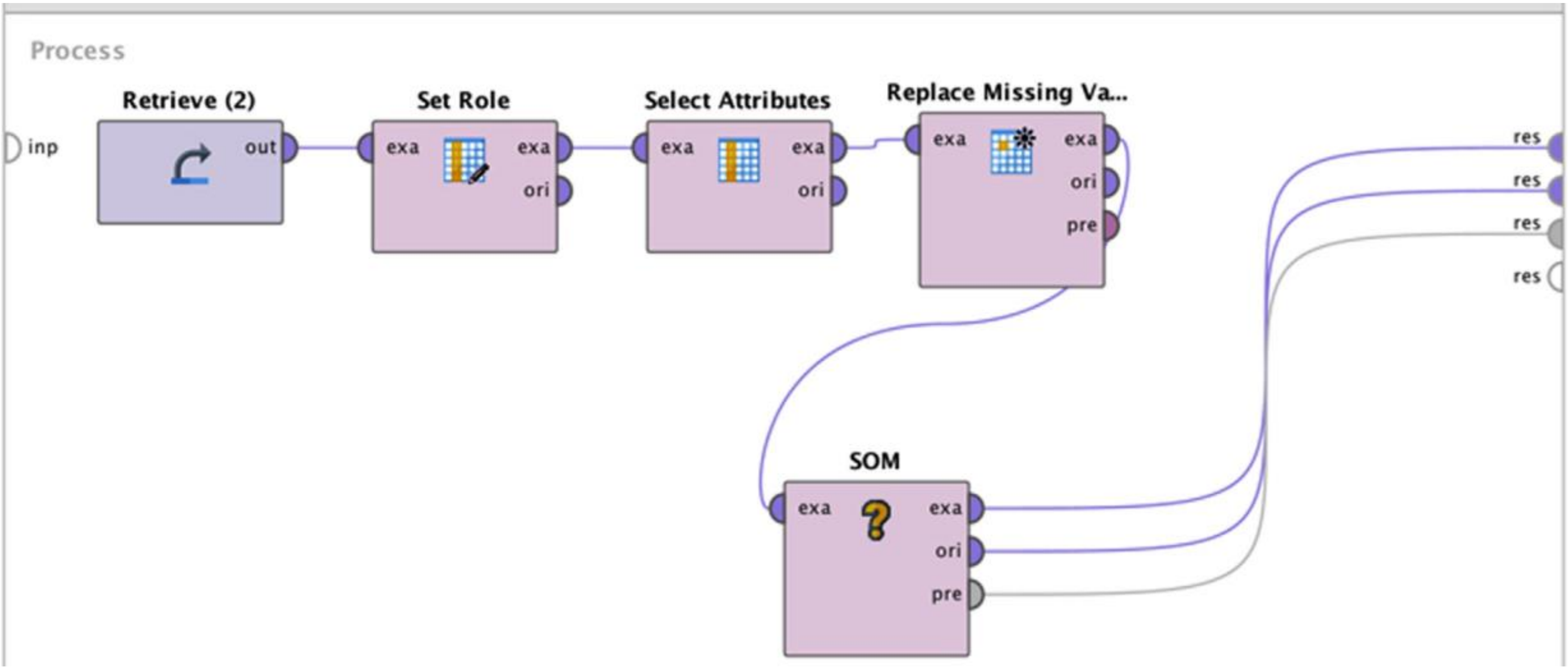
Algorithm

- Step 5: Termination
  - The entire algorithm is continued until no significant centroid updates take place in each run or until the specified number of run count is reached
  - a SOM tends to converge to a solution in most cases but doesn't guarantee an optimal solution
- Step 6: Mapping a New Data Object
  - any new data object can be quickly given a location on the grid space, based on its proximity to the centroids.
  - The characteristics of new data objects can be further understood by studying the neighbors.

SOM in RapidMiner



SOM in RapidMiner

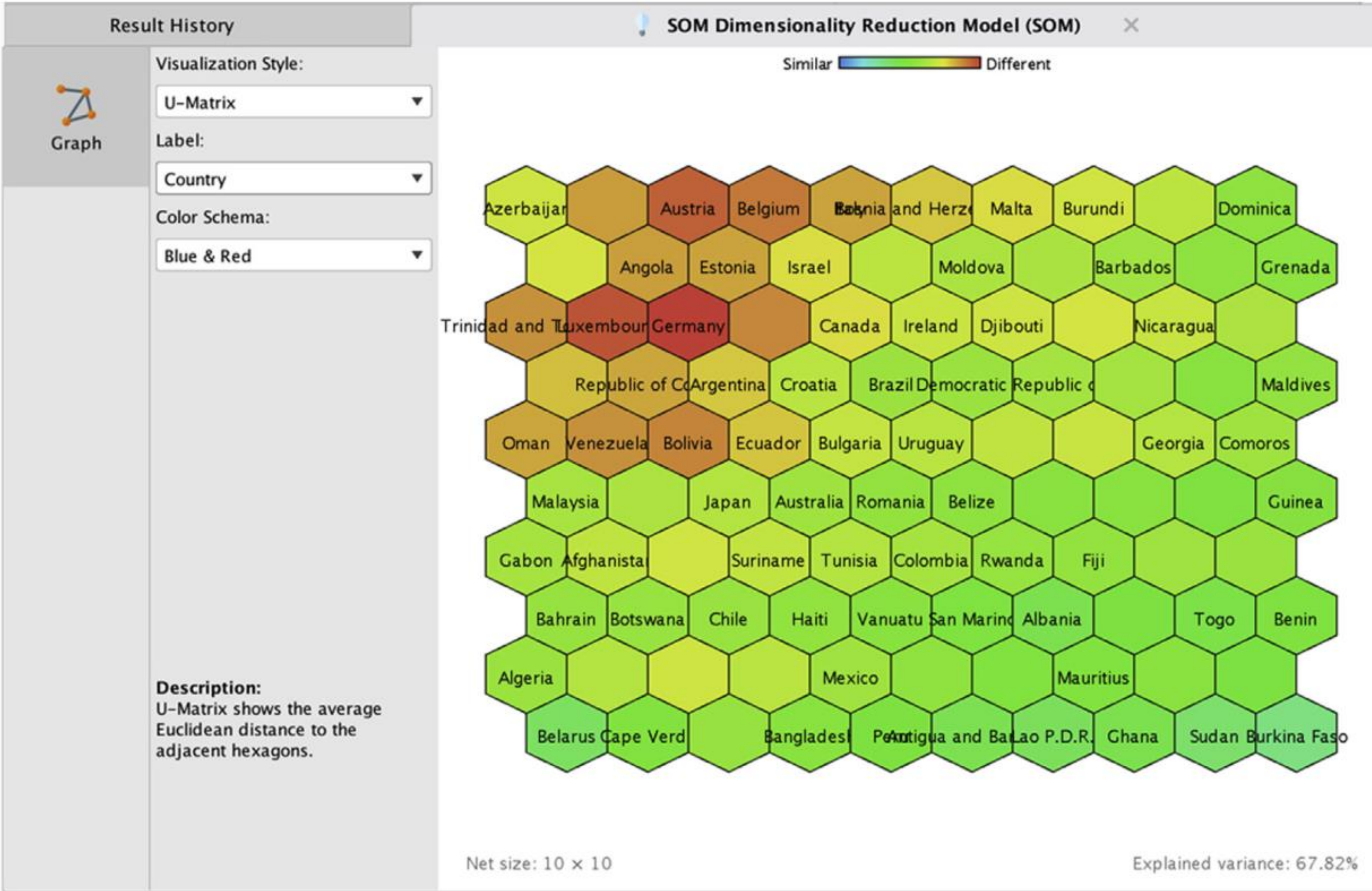


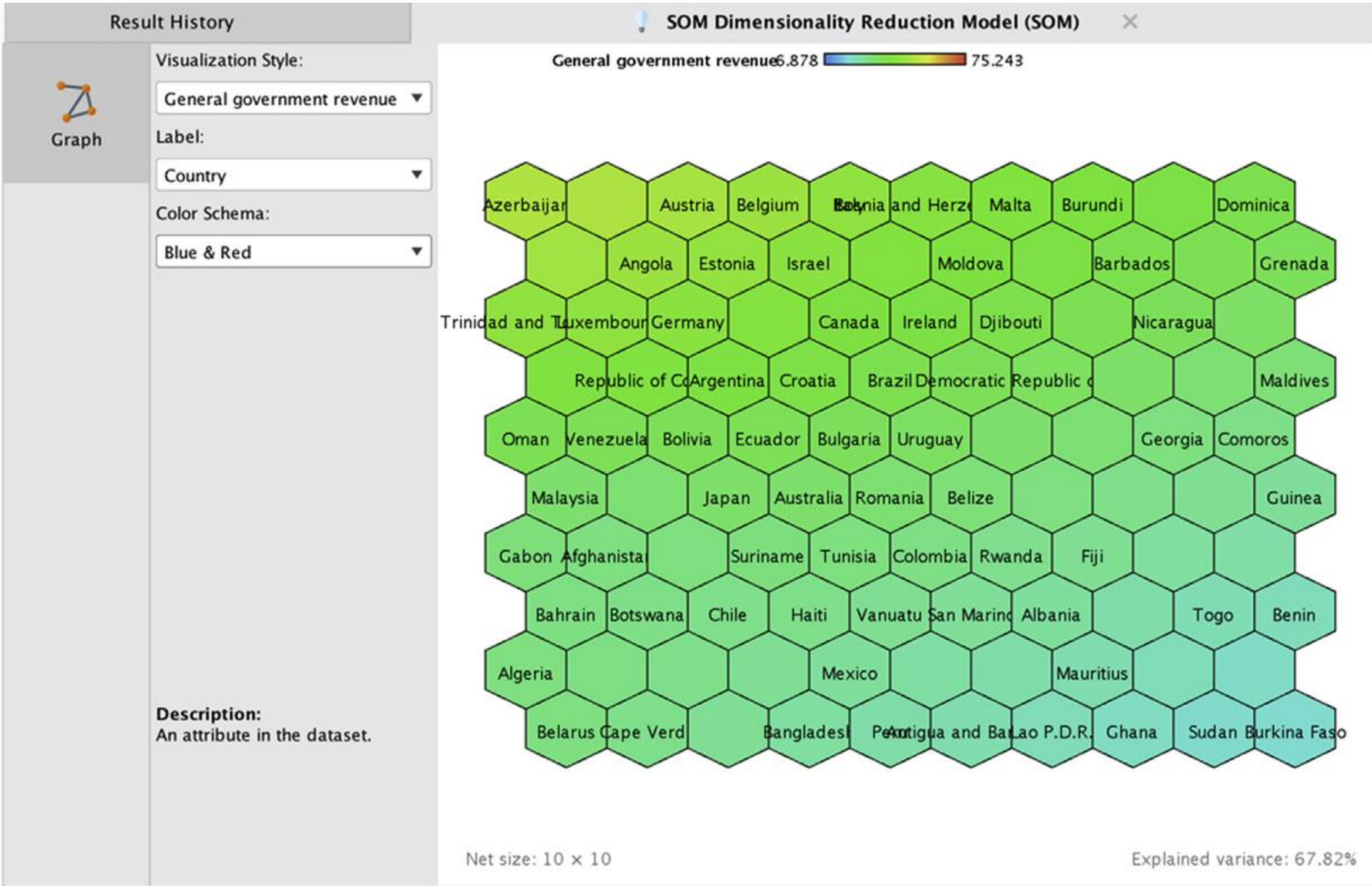
Example

The dataset has 186 records, one for each country, and four attributes in percentage of GDP:

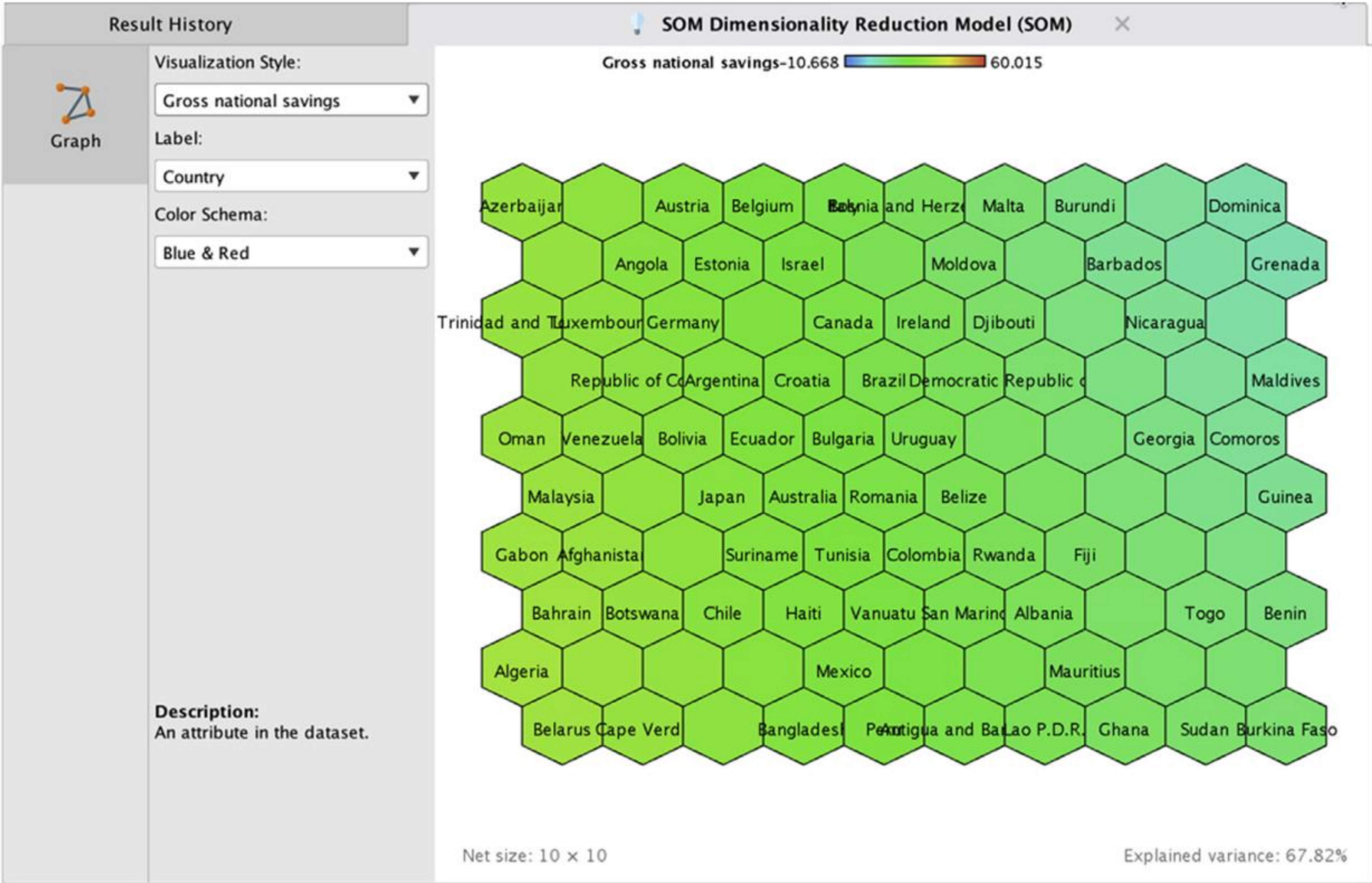
- relative GDP invested
- relative GDP saved
- government revenue, and
- current account balance

Row No.	Country	Current account balance	General government revenue	Gross national savings	Total investment
1	Afghanistan	3.877	21.977	30.398	26.521
2	Albania	−11.372	25.835	14.509	25.886
3	Algeria	7.489	36.458	48.947	41.428
4	Angola	9.024	43.479	21.692	12.668
5	Antigua and...	−13.109	22.430	16.194	29.303
6	Argentina	0.658	37.199	22.595	24.451
7	Armenia	−14.653	20.970	16.660	31.313
8	Australia	−2.870	31.846	23.925	26.794
9	Austria	3.009	48.105	24.611	21.602
10	Azerbaijan	28.423	45.652	46.955	18.532


Result History
SOM Dimensionality Reduction Model (SOM)
Graph
Visualization Style:
U-Matrix
Label:
Country
Color Schema:
Blue & Red
Description:
U-Matrix shows the average Euclidean distance to the adjacent hexagons.
Similar
Different
Azerbaijan
Austria
Belgium
Bosnia and Herze
Malta
Burundi
Dominica
Angola
Estonia
Israel
Moldova
Barbados
Grenada
Trinidad and T
Luxembour
Germany
Canada
Ireland
Djibouti
Nicaragua
Republic of C
Argentina
Croatia
Brazil
Democratic Republic
Maldives
Oman
Venezuela
Bolivia
Ecuador
Bulgaria
Uruguay
Georgia
Comoros
Malaysia
Japan
Australia
Romania
Belize
Guinea
Gabon
Afghanista
Suriname
Tunisia
Colombia
Rwanda
Fiji
Bahrain
Botswana
Chile
Haiti
Vanuatu
San Marino
Albania
Togo
Benin
Algeria
Mexico
Mauritius
Belarus
Cape Verd
Banglades
Antigua and Ba
Lao P.D.R.
Ghana
Sudan
Burkina Faso
Net size: 10 × 10
Explained variance: 67.82%



Result History
SOM Dimensionality Reduction Model (SOM)
Graph
Visualization Style:
General government revenue
Label:
Country
Color Schema:
Blue & Red
Description:
An attribute in the dataset.
General government revenue 6.878
75.243
Azerbaijan
Austria
Belgium
Bosnia and Herze
Malta
Burundi
Dominica
Angola
Estonia
Israel
Moldova
Barbados
Grenada
Trinidad and T
Luxembour
Germany
Canada
Ireland
Djibouti
Nicaragua
Republic of C
Argentina
Croatia
Brazil
Democratic Republic
Maldives
Oman
Venezuela
Bolivia
Ecuador
Bulgaria
Uruguay
Georgia
Comoros
Malaysia
Japan
Australia
Romania
Belize
Guinea
Gabon
Afghanista
Suriname
Tunisia
Colombia
Rwanda
Fiji
Bahrain
Botswana
Chile
Haiti
Vanuatu
San Marino
Albania
Togo
Benin
Algeria
Mexico
Mauritius
Belarus
Cape Verd
Banglades
Antigua and Ba
Lao P.D.R.
Ghana
Sudan
Burkina Faso
Net size: 10 × 10
Explained variance: 67.82%



Result History
SOM Dimensionality Reduction Model (SOM)
Graph
Visualization Style:
Gross national savings
Label:
Country
Color Schema:
Blue & Red
Description:
An attribute in the dataset.
Gross national savings-10.668
60.015
Azerbaijan
Austria
Belgium
Malta
Burundi
Dominica
Angola
Estonia
Israel
Moldova
Barbados
Grenada
Germany
Canada
Ireland
Djibouti
Nicaragua
Argentina
Croatia
Brazil
Maldives
Oman
Venezuela
Bolivia
Ecuador
Bulgaria
Uruguay
Georgia
Comoros
Malaysia
Japan
Australia
Romania
Belize
Guinea
Gabon
Suriname
Tunisia
Colombia
Rwanda
Fiji
Bahrain
Botswana
Chile
Haiti
Vanuatu
Albania
Togo
Benin
Algeria
Mexico
Mauritius
Belarus
Lao P.D.R.
Ghana
Sudan
Burkina Faso
Net size: 10 × 10
Explained variance: 67.82%