







Khiops Covisualization Guide 10.2

-  ■ **Khiops**
 - Optimal data preparation based on discretization and value grouping
 - Scoring models for classification and regression
 - Correlation analysis between pairs of variables
 - Automatic variable construction for multi-table relational mining
-  ■ **Khiops Visualization**
 - Analysis of Khiops results using an interactive visualization tool
-  ■ **Khiops Coclustering**
 - Correlation analysis of two or more variables using a hierarchical coclustering model
 - Prediction models for cluster assignment
-  ■ **Khiops Covisualization**
 - Exploratory analysis of Khiops Coclustering results using an interactive visualization tool

This guide is about the Khiops Covisualization component.

Abstract

Khiops Covisualization is a tool for exploring, interpreting and annotating a hierarchical coclustering model. This tool is complementary to Khiops Coclustering.

It does not require any statistical knowledge from the user.

Summary

| | |
|---|-----------|
| 1. Presentation | 3 |
| 2. Installation | 3 |
| 3. First steps..... | 4 |
| 3.1. Managing views | 5 |
| 3.2. Screenshot | 7 |
| 3.3. Table display | 7 |
| 3.4. Zoom | 9 |
| 4. The different views and features | 9 |
| 4.1. Hierarchy..... | 9 |
| 4.1.1. Renaming the clusters | 10 |
| 4.2. Save Current Hierarchy | 12 |
| 4.3. Composition..... | 13 |
| 4.4. Current Cluster | 15 |
| 4.5. Distribution | 16 |
| 4.6. Dimensions | 17 |
| 4.7. Co-occurrence matrix | 18 |
| 4.7.1. Criteria | 19 |
| 4.7.2. Axis representation | 21 |
| 4.7.3. Contrast | 21 |
| 4.7.4. Summary..... | 22 |
| 4.8. Annotation | 23 |
| 4.9. External data..... | 23 |
| 4.9.1. How to import external data | 23 |
| 4.9.2. Details on the format of external data files | 25 |
| 4.10. Interaction between views | 26 |
| 5. Managing three dimensions or more | 28 |
| 6. Technical limits | 30 |

1. Presentation

Khiops Covisualization is a tool for exploring, interpreting and annotating a hierarchical coclustering model. This tool is complementary to Khiops Coclustering. It does not require any statistical knowledge from the user.

Khiops Covisualization is intended to be integrated into an analytical process involving the Khiops tool upstream and downstream. Upstream, **Khiops Coclustering** produces a coclustering model from a data sample in a back-end process. Downstream, **Khiops Coclustering** allows deploying the synthetic model on new data, at the granularity level chosen by the domain expert using **Khiops Covisualization**.

For more information on the Khiops Coclustering tool, it is strongly advised to refer to the Khiops Coclustering Guide.

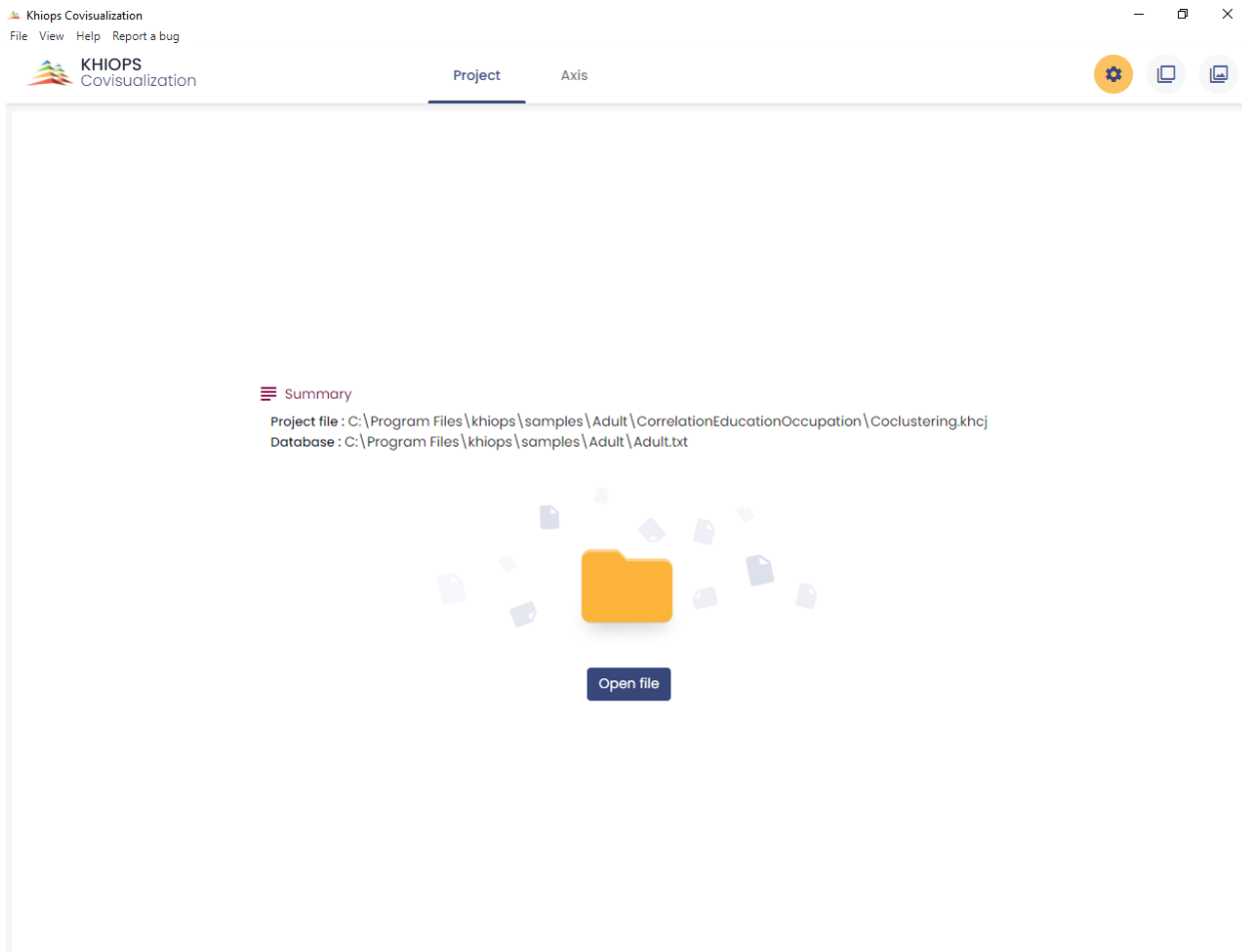
2. Installation

Installation of Khiops Covisualization is supported by the installation of the **Khiops** tool available on the website <https://khiops.org>.

3. First steps

The entry point of Khiops Covisualization is the khcj file generated by Khiops Coclustering at the end of the analysis. A double-click on this file opens Khiops Covisualization.

Khiops Covisualization is composed of several tabbed panes. The first one is the project pane : it presents the report file and database locations.




The second pane is the axis pane : it presents the coclustering.

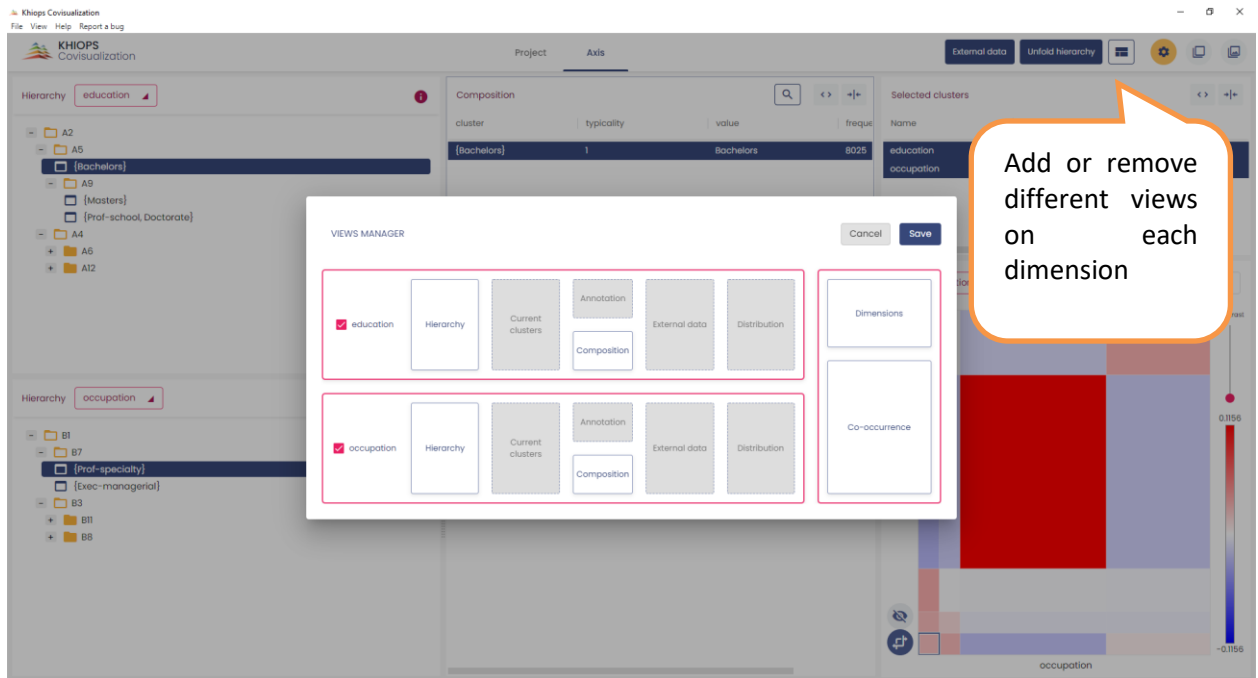
There is a third pane only with three (or more) dimensions : the two "main" dimensions are presented in the second pane as if there were only two dimensions. The other dimensions are contextual, you can view them one by one in the third pane called the context pane.

The coclustering window in the second pane is split into three main parts; there is one part for each variable (or dimension). The right part shows a co-occurrence matrix.



3.1. Managing views

By clicking on the  button, you can add or remove different views on each dimension.

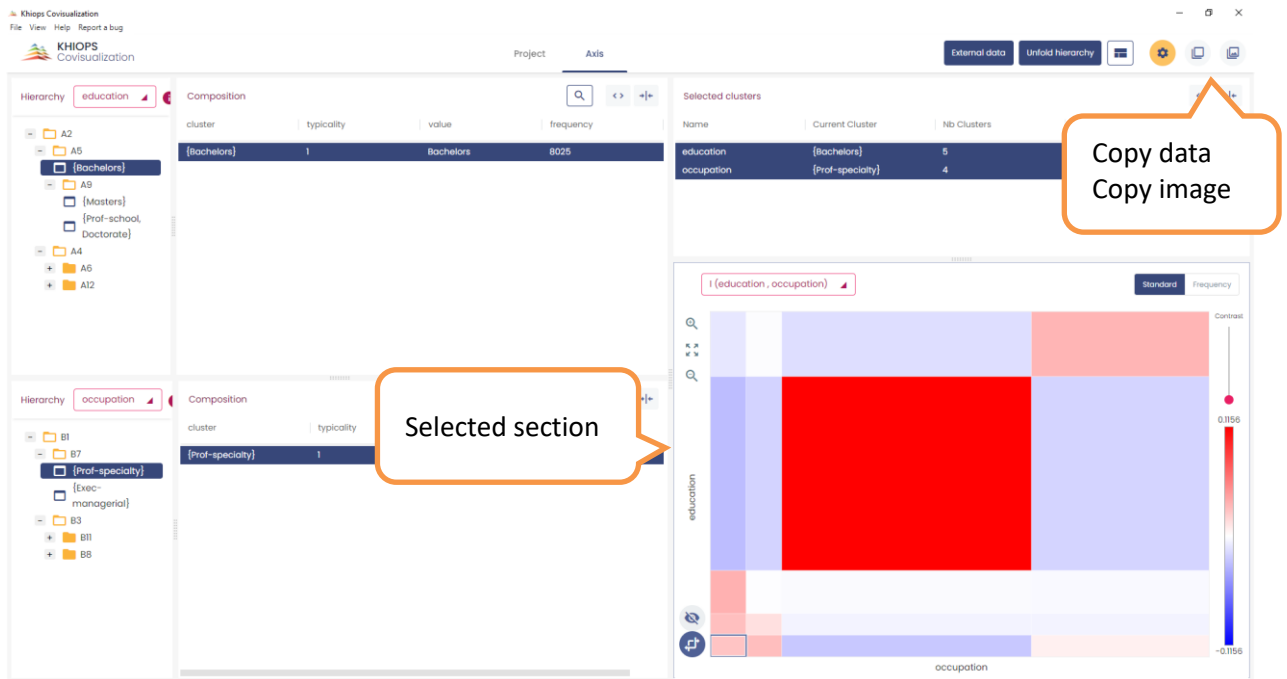



The views manager window reflects the Khiops Covisualization interface. The shaded boxes represent the hidden panels while the white boxes represent the visible panels. You can hide or display a panel by clicking the corresponding checkbox. In the same way you can display only one dimension by clicking the checkboxes before the name of each dimension.


3.2. Screenshot

A useful feature available on all window sections is the screenshot. You can copy to clipboard a selected section by a simple mouse click on one of the top buttons.

A section is selected when it is surrounded by blue lines.




Using  (or [Ctrl-C]) you get a picture in bitmap format that you can import in any image editing tool.

Using  (or [Ctrl-D]) you get the raw data (a table in CSV format), that you can import in any text editing tool.

Click on a column label to sort the column's values in one of three ways: ascending, descending or initial order.

3.3. Table display

Clicking  opens a list of columns you can hide or display. The red dot indicates that at least one column is hidden

Current clusters

| Name | Size | Frequency | Interest |
|--------------------|------|-----------|----------|
| {Bachelors} | 1 | 8025 | 1 |
| {Masters} | 1 | 2657 | 0.93295 |
| {Prof-school, C | 2 | 1428 | 0.88103 |
| {HS-grad} | 1 | 15784 | 0.98451 |
| {7th-8th, 9th, E | 4 | 2550 | 0.78203 |
| {11th, 10th, 12th} | 3 | 3858 | 0.73857 |
| {Some-college} | 1 | 10878 | 0.67075 |
| {Assoc-voc} | 1 | 2061 | 0.38201 |
| {Assoc-acdm} | 1 | 1601 | 0.326274 |

Composition

- Name
- Father
- Size
- Frequency
- Interest
- Hierarchical Level
- Rank



allow to automatically redimension the columns of the table.



opens a search box.

Long tables are split into pages of at most 500 lines.

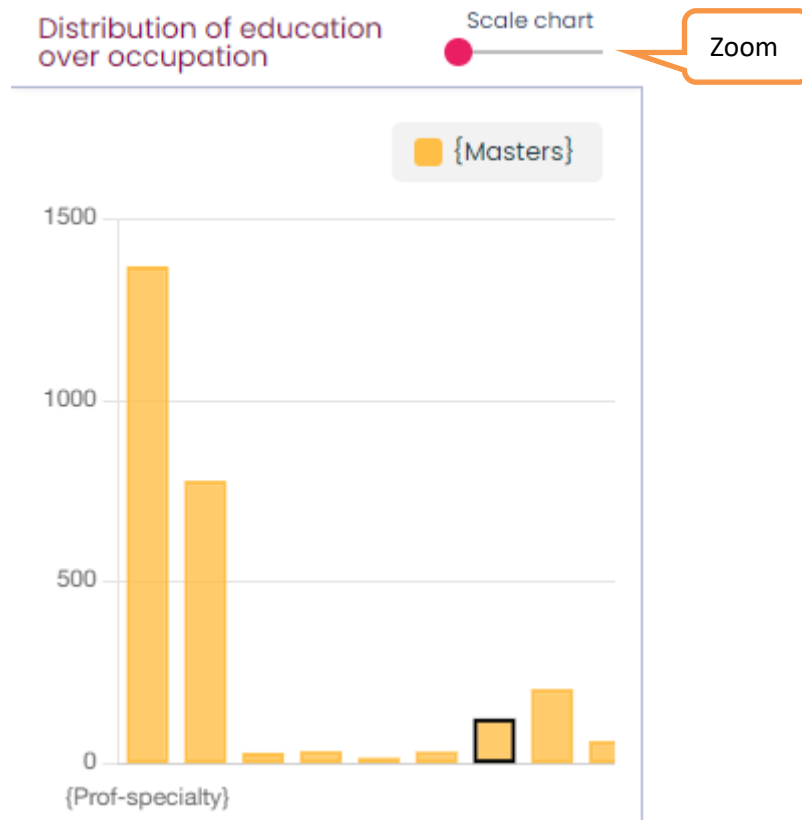
Current clusters

| Name | Size | Frequency | Interest |
|---------------|------|-----------|----------|
| {158, 6233, 4 | 4 | 16607 | 0.874391 |
| {12462, 173 | 4 | 11252 | 0.786903 |
| {3633, 103 | 4 | 9213 | 0.667522 |
| {11661, 822 | 4 | 10966 | 0.784932 |
| {5011, 1575 | 4 | 12024 | 0.784615 |
| {11227, 156 | 4 | 11125 | 0.685951 |
| {11348, 161 | 4 | 9829 | 0.749308 |
| {16027, 15 | 4 | 6705 | 0.718597 |
| {13587, 23, | 4 | 5389 | 0.950277 |
| {7034, 110 | 4 | 4826 | 0.698162 |
| {18545, 13 | 4 | 5290 | 0.641812 |

1 to 100 of 204 | Page 1 of 3

3.4. Zoom

The sliding bars allow zooming on the chart.

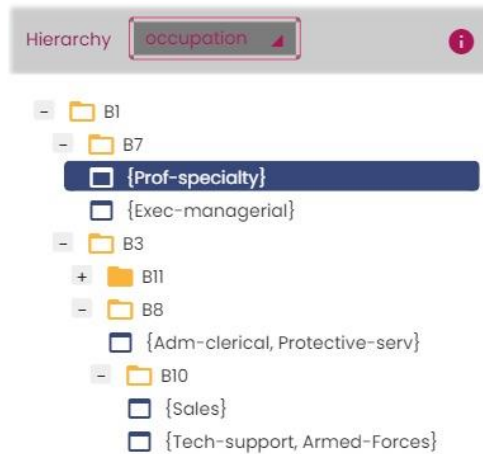


4. The different views and features




We will now explore the various features offered by Khiops Covisualization using a two dimensional example. Khiops covisualization allows to explore coclustering of more than 2 dimensions: these features are detailed in the section "5. Managing three dimensions or more".

4.1. Hierarchy

This view shows the cluster hierarchy. It allows to navigate and to modify the hierarchy of clusters. The hierarchy is presented like a file manager, each directory being a cluster. By opening or closing a cluster, you can choose the appropriate unfolding of the hierarchy.



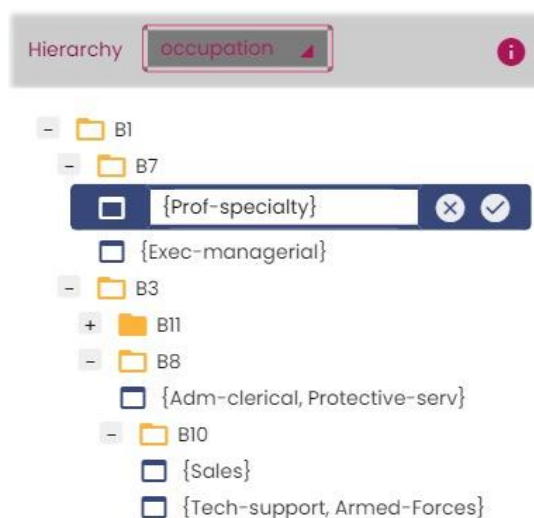
A cluster is represented in three ways:

-  An unfolded cluster: it contains clusters that are visible in this unfolding. You can close this cluster.
-  A folded cluster: it contains clusters that are not visible in this unfolding. You can open this cluster.
-  A “terminal” cluster: it does not contain any clusters. You can neither open nor close it.

Navigation, selection, folding and unfolding can be done via the arrow keys.

4.1.1. Renaming the clusters

Clusters can be renamed by double-clicking on the cluster name. To restore its original name, you have to rename it with an empty string.



4.1.1.1. Unfold Hierarchy

The best way to fold or unfold hierarchies is to use the “Unfold hierarchy” view. The button “Unfold hierarchy” on the top of the user interface opens a new window.

This window allows choosing the best number of clusters on both dimensions: optimal unfolding of each partition so as to keep the most informative model. The top chart shows the evolution of the level according to the number of clusters, that is the percentage of information kept at each granularity of the hierarchy. This allows obtaining a simplified and easily interpretable hierarchy with few clusters while keeping most of the information. In real cases, 80% of the information often comes from a small fraction of the clusters.

The bottom chart shows the detailed number of clusters for each dimension (a mouse-click on the legend highlights the corresponding curve).

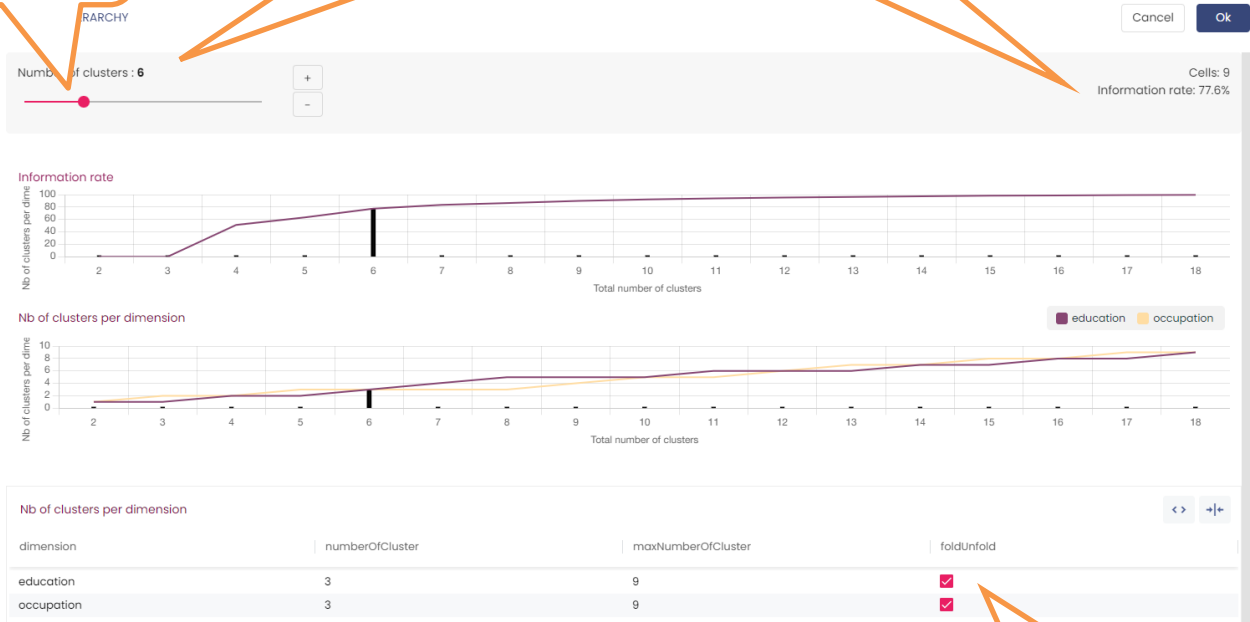
Once the number of clusters is chosen (and therefore the information kept with the coclustering), by clicking the button “Ok”, you return to the main window where the hierarchies are now unfolded.

The checkboxes on the bottom of this view allow selecting which hierarchy to unfold. By default all hierarchies are unfolded.

Slide bar allowing choosing the optimal number of clusters

Total number of clusters for the current level of hierarchy

Percentage of information kept with this current level of hierarchy



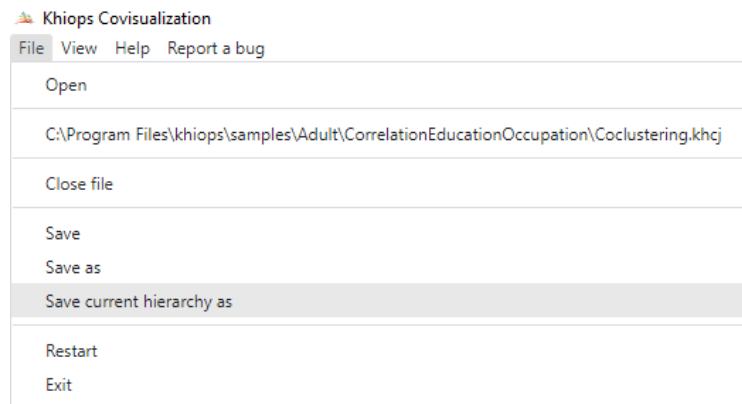
Number of clusters on each dimension for the current level of hierarchy



Maximum number of clusters of each dimension

Apply or not the unfolding

4.2. Save Current Hierarchy

Once you have chosen the desired granularity of the coclustering, you can generate a simplified coclustering corresponding to the current unfolding via the menu “Save current hierarchy...”.



The resulting new file contains only the clusters that are visible in the current unfolding (all folded clusters represented with  become terminal cluster ).

This is useful to share a simpler coclustering model with other users. In addition, this new model can be deployed by the Khiops Coclustering tool (see the Khiops Coclustering tool guide for more information on the process).

4.3. Composition

Given a selected cluster in the hierarchy view, the composition view presents in a table all value items (value in the dimension) which are grouped in this cluster. An item belongs to a cluster if it belongs to one of its sub-cluster.

Composition

| cluster | terminalCluster | rank | typicality | value | frequency |
|-----------------------|-----------------------|------|------------|-----------------|-----------|
| {Prof-specialty, Arr} | {Prof-specialty, Arr} | 1 | 1 | Prof-specialty | 5787 |
| {Prof-specialty, Arr} | {Prof-specialty, Arr} | 1 | 0.0027 | Armed-Forces | 13 |
| {Exec-manager, Arr} | {Exec-managerial} | 3 | 1 | Exec-manageric | 3973 |
| {Machine-op-in, Arr} | {Machine-op-insp} | 5 | 1 | Machine-op-ins | 1852 |
| {Machine-op-in, Arr} | {Machine-op-insp} | 5 | 0.9132 | Transport-movir | 1578 |
| {Machine-op-in, Arr} | {Machine-op-insp} | 5 | 0.7806 | Handlers-clean€ | 1259 |
| {Craft-repair} | {Craft-repair} | 7 | 1 | Craft-repair | 3937 |
| {Other-service} | {Other-service} | 9 | 1 | Other-service | 2977 |
| {Farming-fishin, P} | {Farming-fishing, P} | 11 | 1 | Farming-fishing | 931 |
| {Farming-fishin, P} | {Farming-fishing, P} | 11 | 0.2425 | Priv-house-serv | 116 |
| {Adm-clerical, F} | {Adm-clerical, Prot} | 13 | 1 | Adm-clerical | 3637 |
| {Adm-clerical, F} | {Adm-clerical, Prot} | 13 | 0.2572 | Protective-serv | 643 |
| {Sales} | {Sales} | 15 | 1 | Sales | 3627 |
| {Tech-support} | {Tech-support} | 17 | 1 | Tech-support | 937 |

This table has one line per item, the value of which is displayed in the column “Value”. The column “Cluster” contains the cluster that contains the item in the current unfolding.

The column “Terminal Cluster” contains the smallest cluster in the hierarchy that contains the item. This is not necessarily a cluster visible in the hierarchy view unless the hierarchy is completely unfolded.

The column “Rank” provides the ordering of the clusters in the hierarchy view. A cluster with a rank value of 1 is the first one visible in the hierarchy.

The “Typicality” of an item is a value between 0 and 1: items with typicality close to 1 are the most representative of the cluster, while items with low typicality are the least representative.

The search edit box allows finding values in the selected cluster. You can find all items containing the substring entered in the search edit box. In the example below, the user finds the item with value “Adm-clerical”. In the hierarchy, the cluster “Adm-clerical, Protec ..” is highlighted; this is the cluster that owns this item.



The screenshot displays the 'Hierarchy' view on the left and the 'Composition' table on the right. The hierarchy tree shows a path from A1 to A8, with the last item selected. The composition table shows two rows of data with columns for cluster, terminalCluster, rank, typicality, value, and frequency.

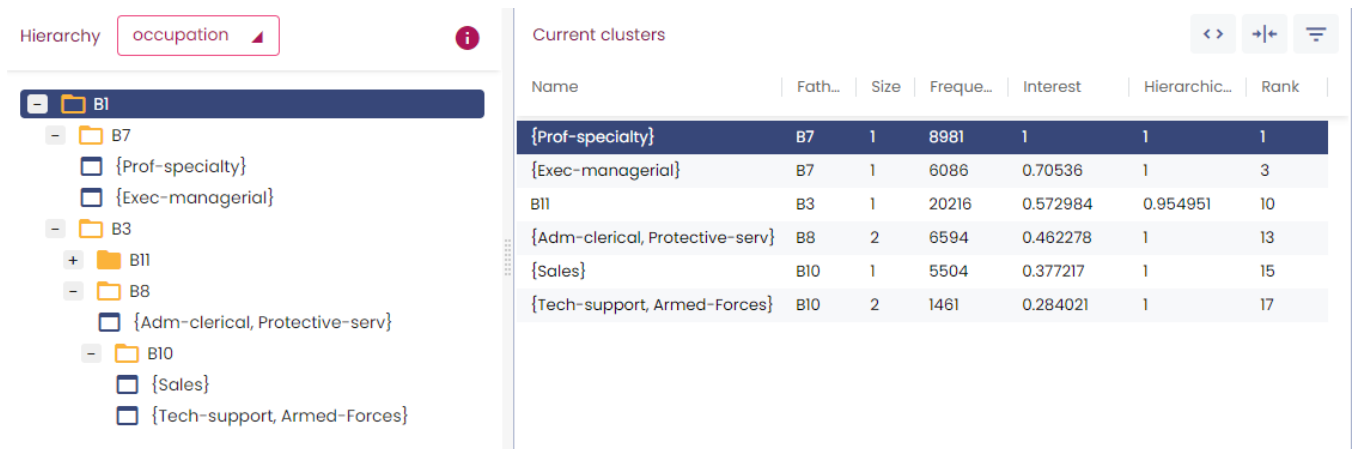
| cluster | terminalCluster | rank | typicality | value | frequency |
|------------------|---------------------|------|------------|-----------------|-----------|
| {Adm-clerical, P | {Adm-clerical, Prot | 13 | 1 | Adm-clerical | 3637 |
| {Adm-clerical, F | {Adm-clerical, Prot | 13 | 0.2572 | Protective-serv | 643 |

A double-click on a line shows the composition of the cluster containing this item. Accordingly, it modifies the selected cluster in the hierarchy view.

4.4. Current Cluster

This table is a flat view of the hierarchy. Each line of the table contains a terminal or a fold cluster:

 or  in the hierarchy view.



The screenshot shows the 'Hierarchy' view on the left with 'occupation' selected. The 'Current clusters' table on the right displays the following data:

| Name | Fath... | Size | Freque... | Interest | Hierarchic... | Rank |
|---------------------------------|---------|------|-----------|----------|---------------|------|
| {Prof-specialty} | B7 | 1 | 8981 | 1 | 1 | 1 |
| {Exec-managerial} | B7 | 1 | 6086 | 0.70536 | 1 | 3 |
| B11 | B3 | 1 | 20216 | 0.572984 | 0.954951 | 10 |
| {Adm-clerical, Protective-serv} | B8 | 2 | 6594 | 0.462278 | 1 | 13 |
| {Sales} | B10 | 1 | 5504 | 0.377217 | 1 | 15 |
| {Tech-support, Armed-Forces} | B10 | 2 | 1461 | 0.284021 | 1 | 17 |

The “Current Cluster” view gives useful informations on clusters:

- Father: the cluster containing this cluster.
- Size: the number of different items that belong to this cluster.
- Frequency: the number of occurrence of these items.
- Interest: a value between 0 and 1. Clusters with interest close to 1 are the most informative.
- Hierarchical level: normalized measure between 0 (all is folded) and 1 (all is unfolded); it represents the information kept for the current unfolding of the hierarchy and can be interpreted as a distance to the root cluster.
- Rank: ordering of the clusters in the hierarchical view.

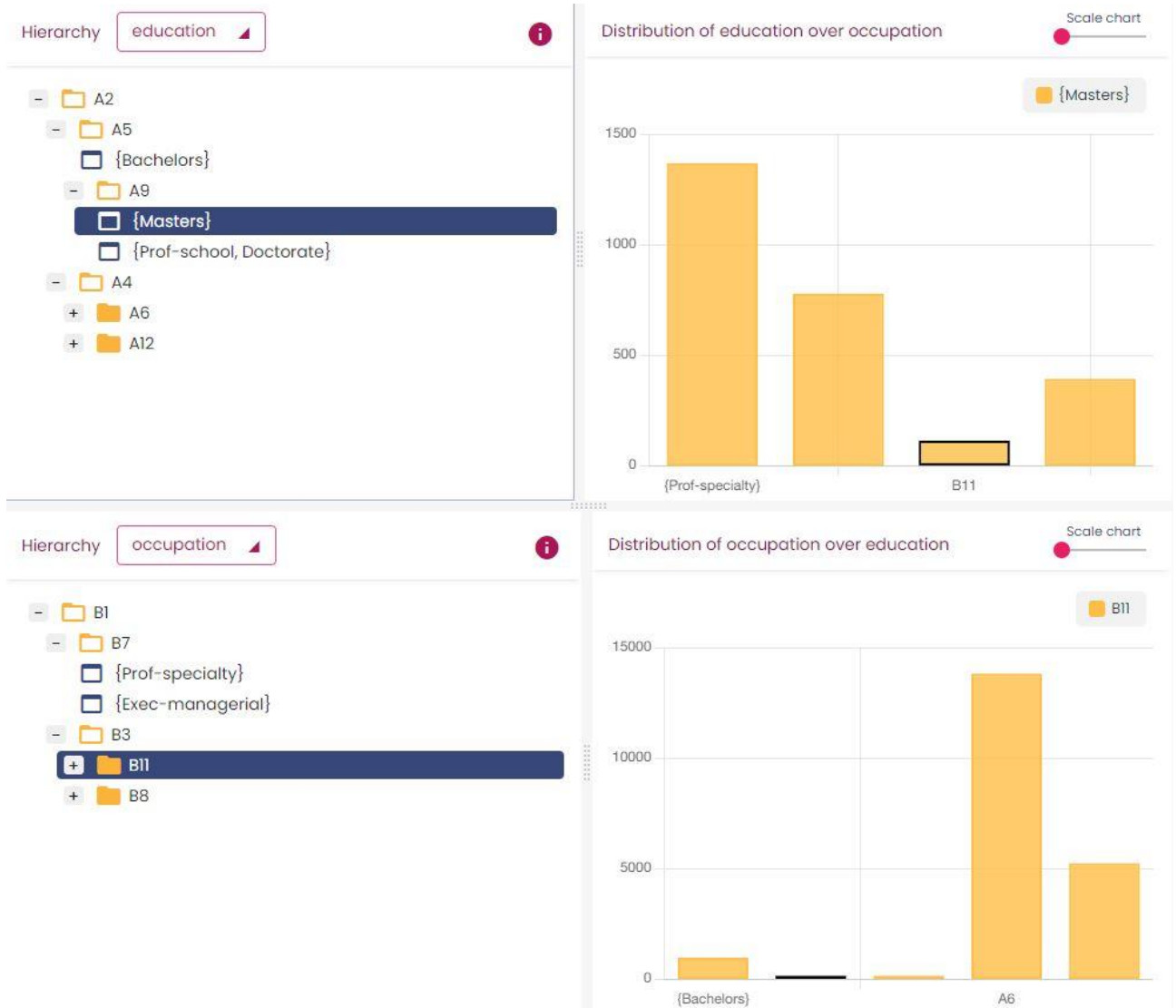
By default, the columns Father, Hierarchical Level and Rank are hidden but they can be displayed using

the  button (cf section 3.3. Table display).

4.5. Distribution

This view allows visualizing the distribution of a dimension on the selected cluster of the other dimension.

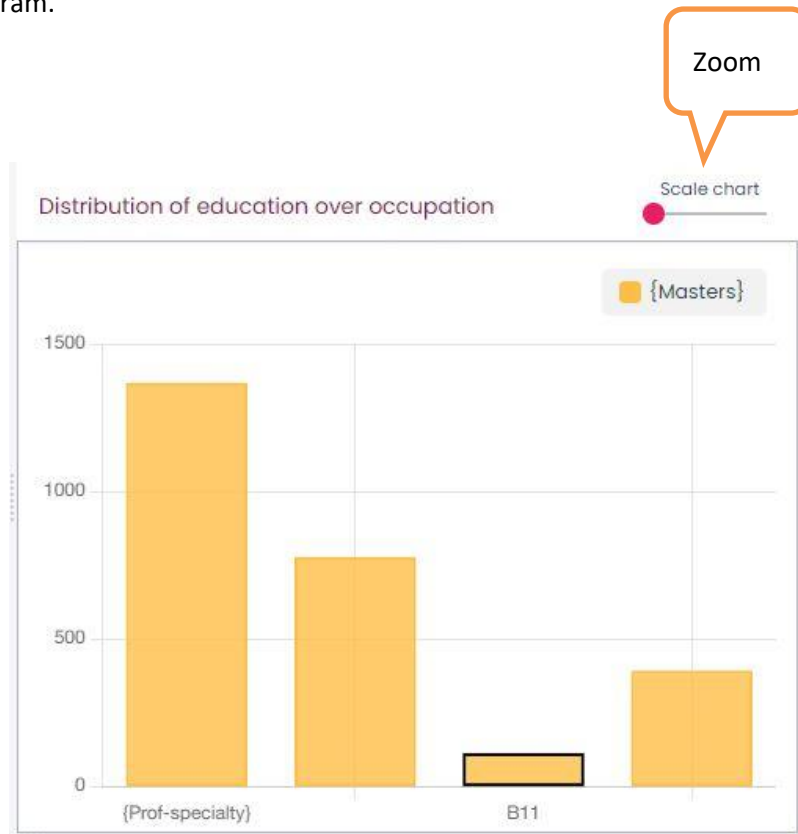
In the example below, the upper histogram represents the distribution of “occupation” for the selected cluster “Masters” of education. The lower histogram represents the distribution of “education” for the selected cluster “B11” of “occupation”.



When a cluster is selected in a hierarchy, its corresponding bar is highlighted in the histogram. You can select another cluster directly on the Distribution in two ways:

- by clicking on a bar of the histogram,
- by navigating in the bar chart with arrow keys

The different views are interactive: when one unfolds a hierarchy, it modifies the distribution view. In the same way, selecting a bar on the chart changes the selected cluster on the hierarchy. A slide-bar allows zooming in the histogram.



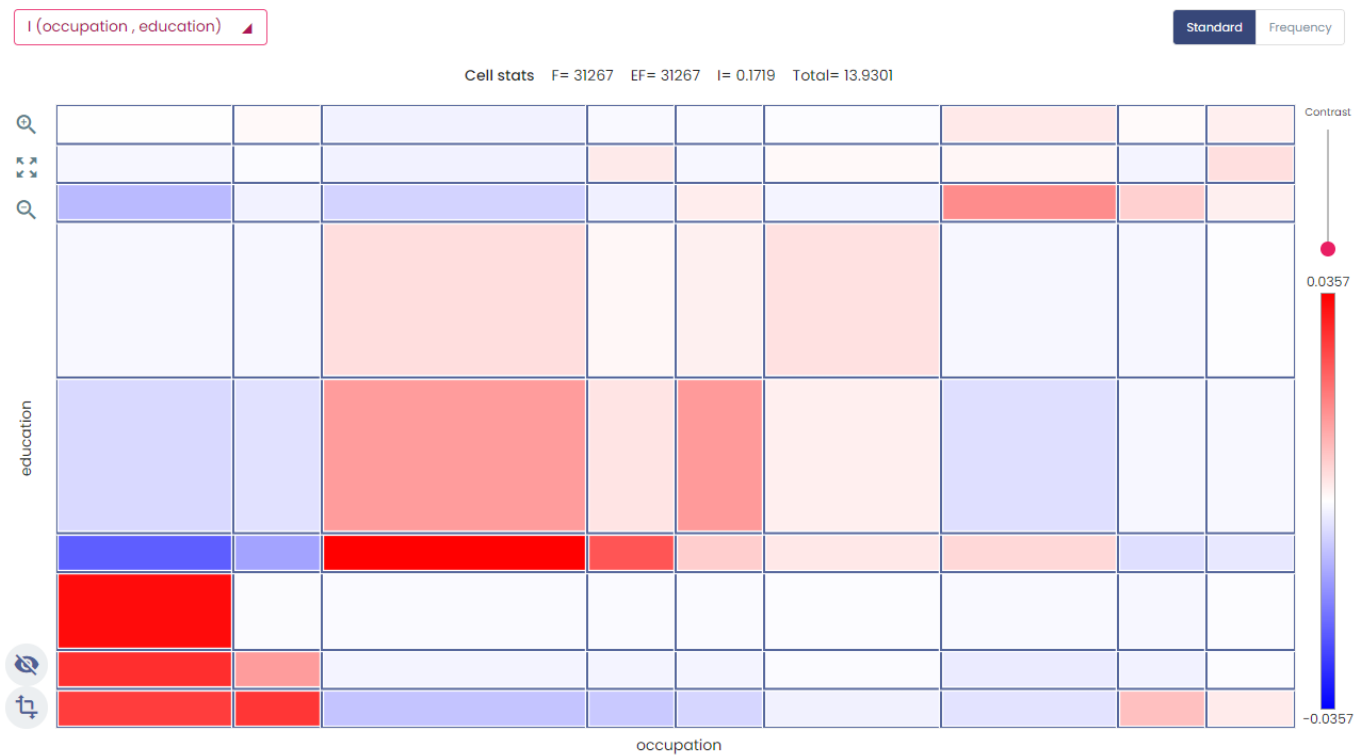
4.6. Dimensions


This view gives a global view of the state of the coclustering at the current level of unfolding: for each dimension it shows the selected cluster and the number of cluster for the current unfolding.

Selected clusters

| Name | Current Cluster | Nb Clusters |
|------------|-----------------|-------------|
| education | A2 | 9 |
| occupation | B11 | 7 |

4.7. Co-occurrence matrix



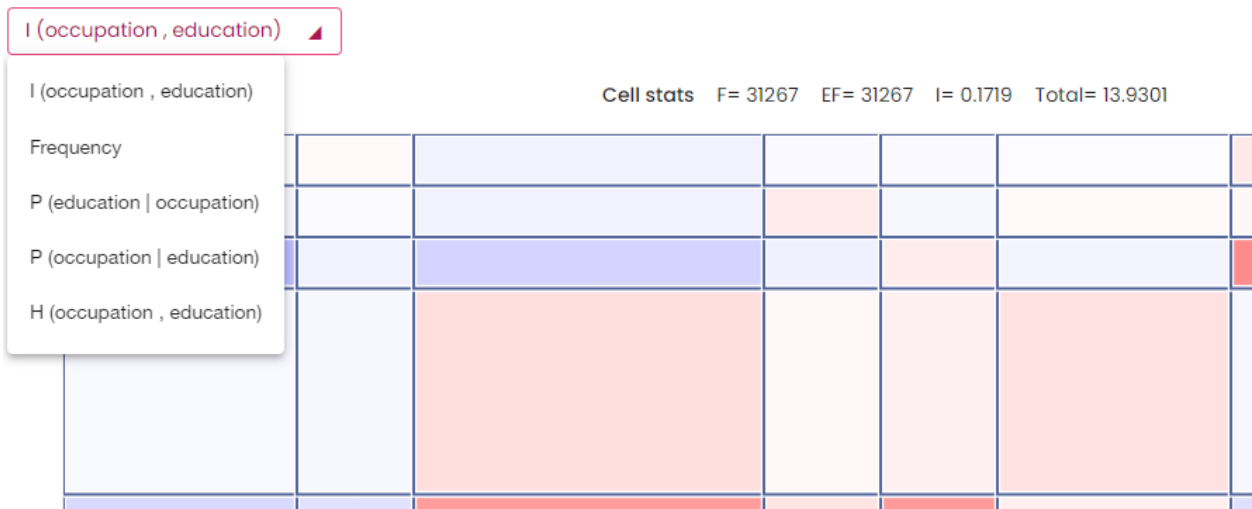
The co-occurrence matrix allows a direct visualization of both partitions jointly. The current hierarchy of each dimension is represented on one chart. The horizontal axis corresponds to the first dimension, the vertical axis to the second one. Axis can be switched using the  button.

Accordingly, the content of the matrix is modified if you fold or unfold the hierarchy.

Selecting a cell in the matrix (by clicking) amounts to select two clusters, one on each hierarchy. And *vice versa*, selecting a cluster in the hierarchy view amounts to select another cell in the matrix. You can zoom in the matrix with the mouse wheel, by click and drag, or by clicking on the '+' and '-' magnifying glass icons.

4.7.1. Criteria

The matrix allows visualizing five criteria. The visualized criterion is chosen by clicking on the combo on the top of the matrix.

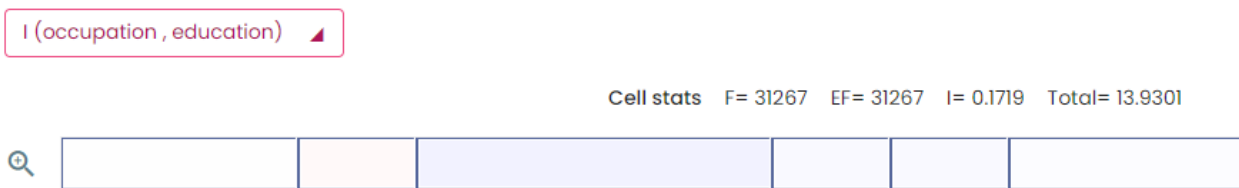


The default criterion is the mutual information (I). In this case, the cells are colored according to their contribution to this criterion:

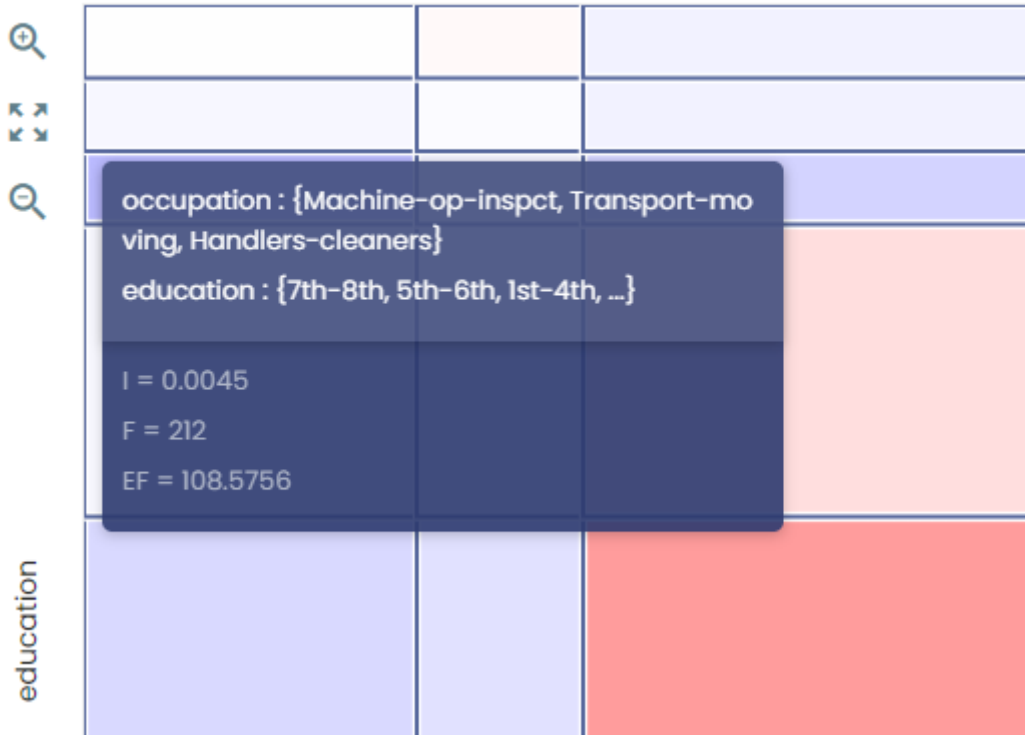
- In red: cells with frequency higher than expected in case of independence.
- In blue: cells with frequency lower than expected in case of independence.

The other criteria are the frequency, the conditional probabilities and the Hellinger distance.

Above the matrix, the value of the selected cell for the current criterion is shown as well as its frequency and its expected frequency in case of independence of the two dimensions (EF). Moreover when the criterion is the mutual information, the total mutual information is displayed.

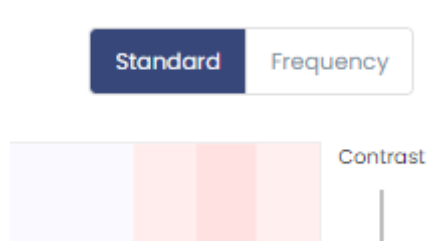


These informations are also displayed on a tooltip by moving the mouse over the cells. Moreover, these tooltips display the clusters that constitute the cell.



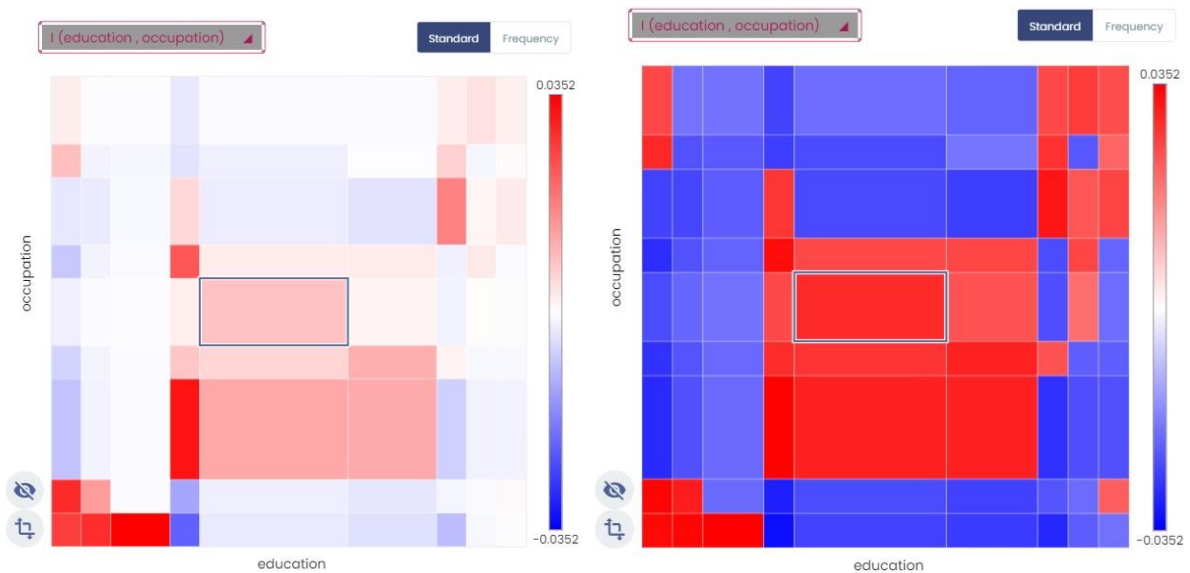
4.7.2. Axis representation

By default, the axis of the matrix represent the cluster size: range of the interval in case of a numerical cluster and number of values in the group in case of a categorical cluster. You can choose another representation with the button below the matrix. By choosing "Frequency", the axis represent the frequency of the clusters.

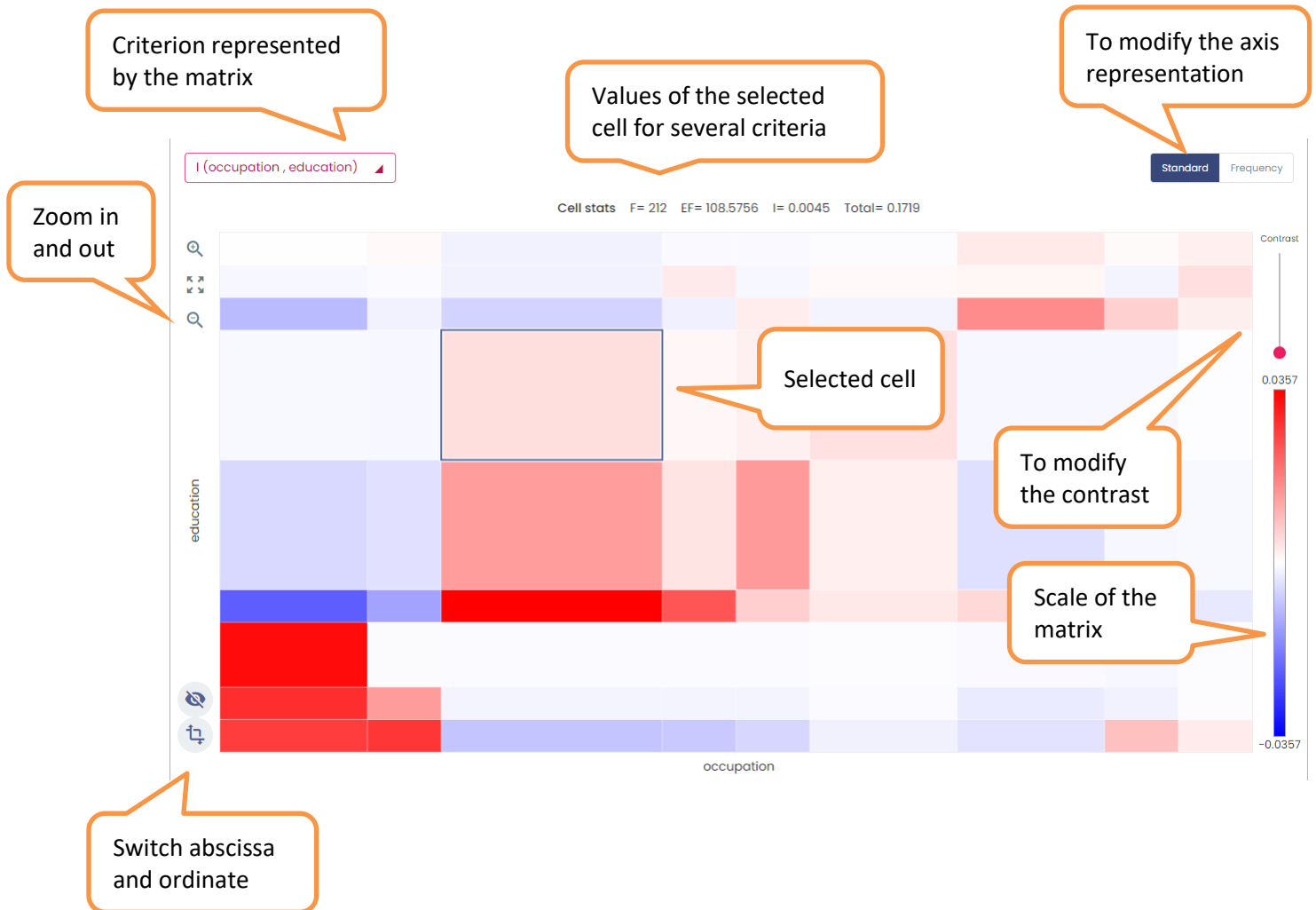


4.7.3. Contrast

The slide bar located at the right of the matrix increases the color contrast in the matrix : the red cells are redder and blue cells are bluer. The picture bellow shows the same matrix with different contrasts.



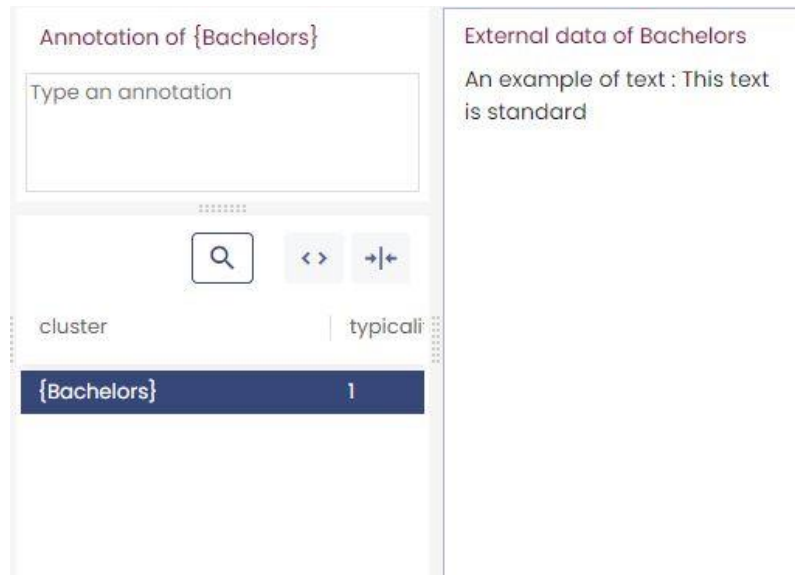
4.7.4. Summary



4.8. Annotation

This view is a simple text box that allows annotating each cluster. The annotation is saved into the khcj file.

4.9. External data



Khiops Covisualization can integrate external data associated to each cluster element. The selection of an item in the composition view displays the corresponding data in the “External data” view. There are 3 types of data fields: textual, numerical and categorical. The display is optimized according to the data type. Several fields can be associated for each item.

External data of Some-college

An example of text
: Here there are
line
breaks...

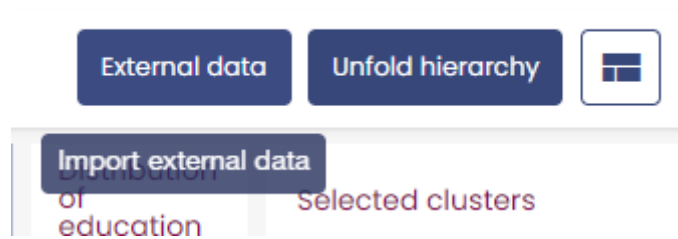
4.9.1. How to import external data

The external data is imported via text files. There is one file for each dimension. The external data files are tab-separated value files; the first line of these files contains the column labels. There is one column for each data field and a “key” column to link data to items.

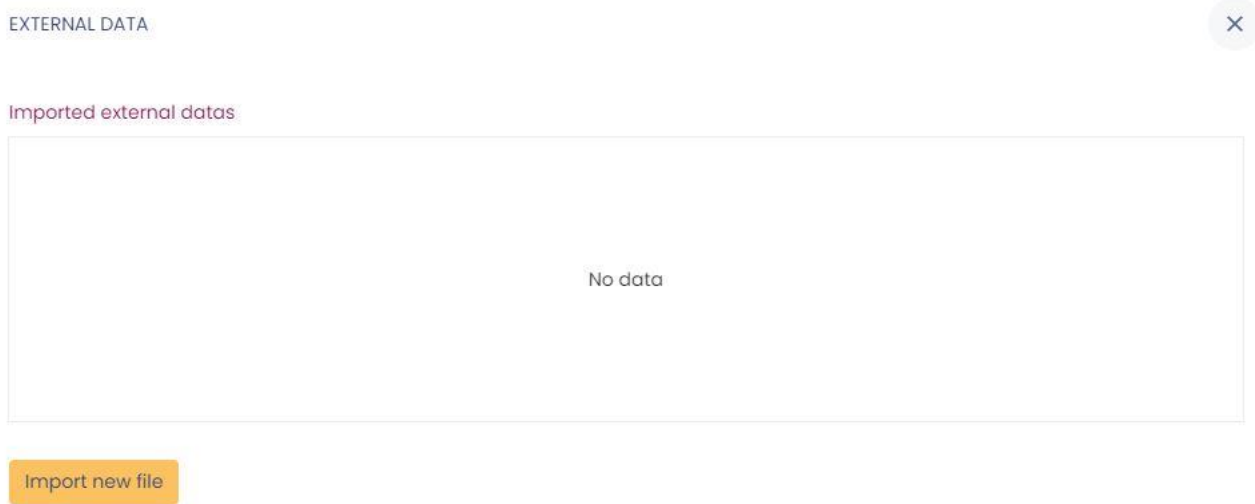
The following table shows a sample of an external data file. It associates three fields to each item: a text, the age and the weight.

| Name | Text | Age | Weight |
|-----------|-----------|-----|--------|
| Bachelors | Some text | 25 | 60 |
| Masters | ... | ... | ... |

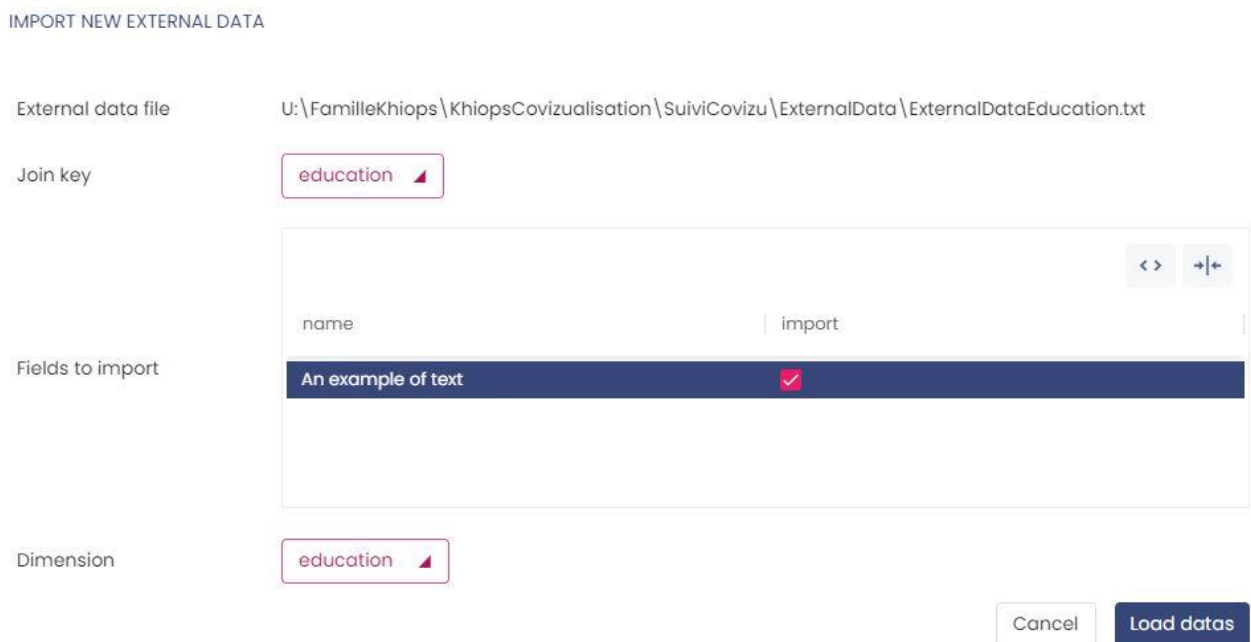
Importing the external files into Khiops Covisualization is performed by clicking on the button “External data...”



A new window appears, to choose the location of the external data file.



Once the file location is filled, you have to choose the join key among the columns of the file and the dimension to which the data belongs. After the import (“Load datas” button), Khiops Covisualization displays these data in the “external data” view.



4.9.2. Details on the format of external data files

As presented above, the external data files are tab-separated value files. The text type has a special format. This format allows you to display multiple lines in a text field for an item.

All characters are allowed, however three characters have a special role:

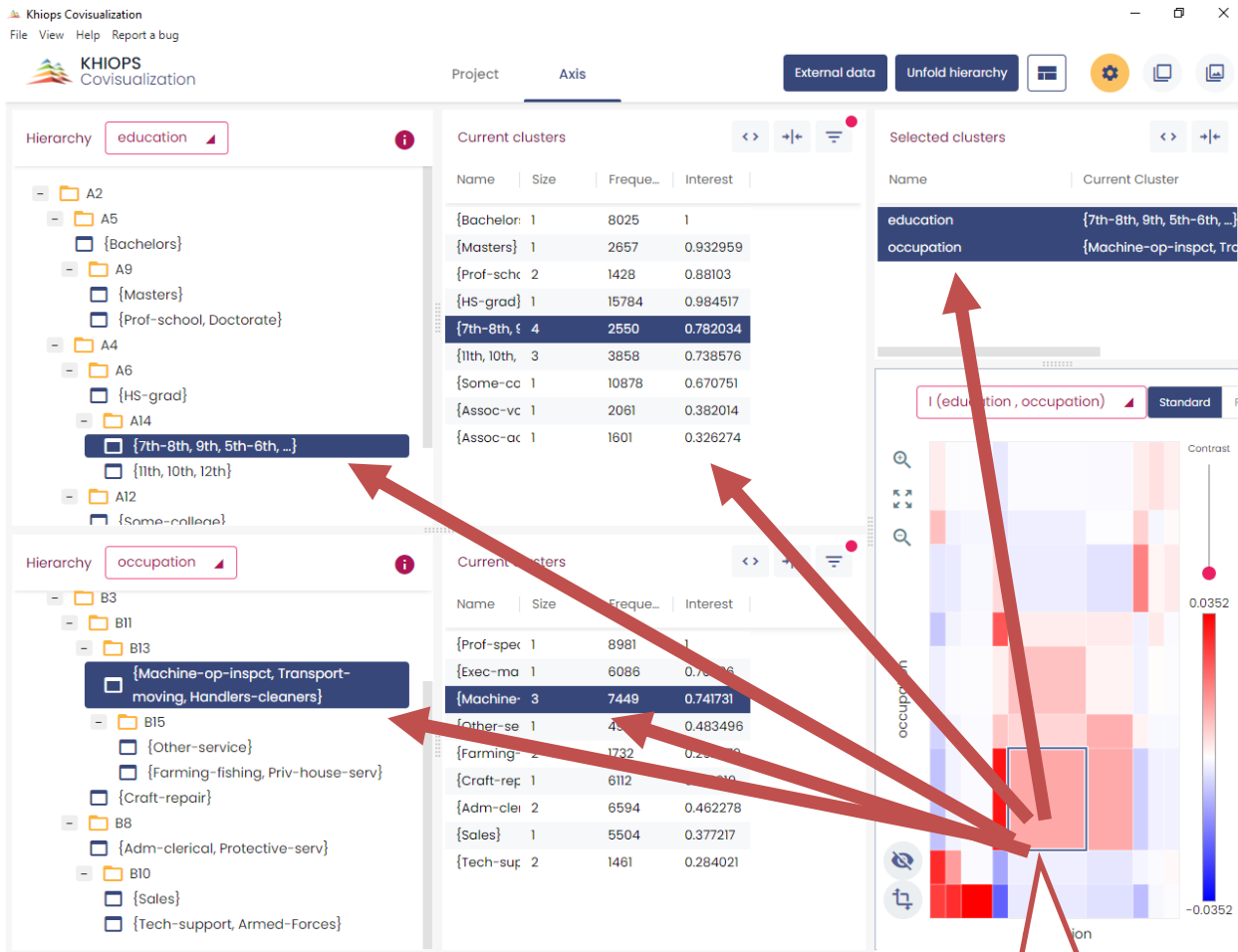
- The field separator : “\t” (tab)
- The new line character : “\n” (eol)
- The beginning or end of field indicator : «"» (double-quote)

If a field contains the tab character and/or the new line character, this field must be surrounded by two double-quotes «"». If a field contains the double-quote character, this character must be doubled. The following table summarizes the different correct ways of writing text fields.

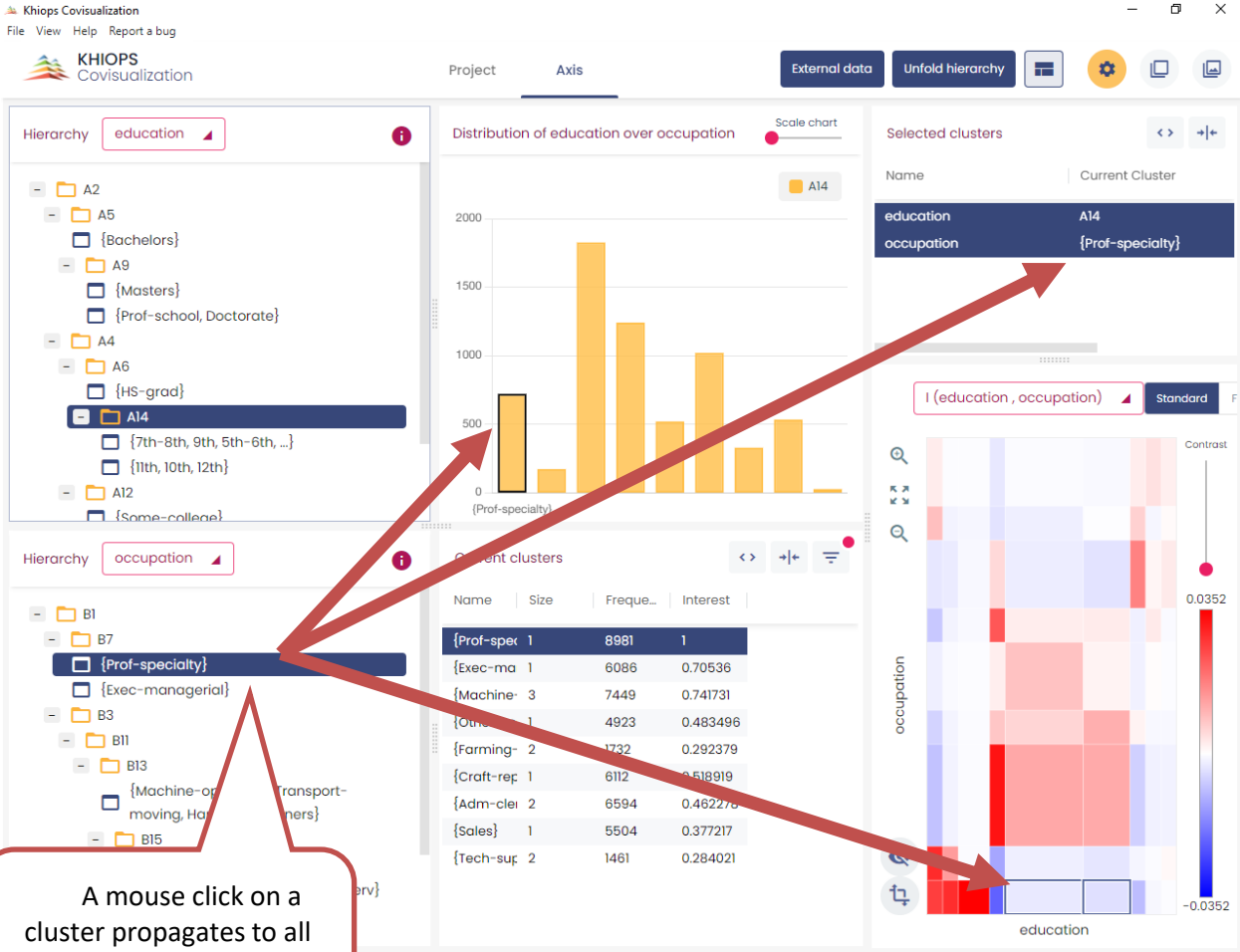
| Key | An example of text |
|--------------|--|
| Bachelors | This text is standard |
| HS-Grad | This one too (;*\ ,.....) |
| Masters | "By cons, it contains a tab" |
| Some-college | "Here there are line breaks..." |
| 10th, 12th | "Here is more complicated, there are ""double quotes""." |

4.10. Interaction between views

All views are interactive : an action on a view often changes several other views. Screenshots below present several interactions.



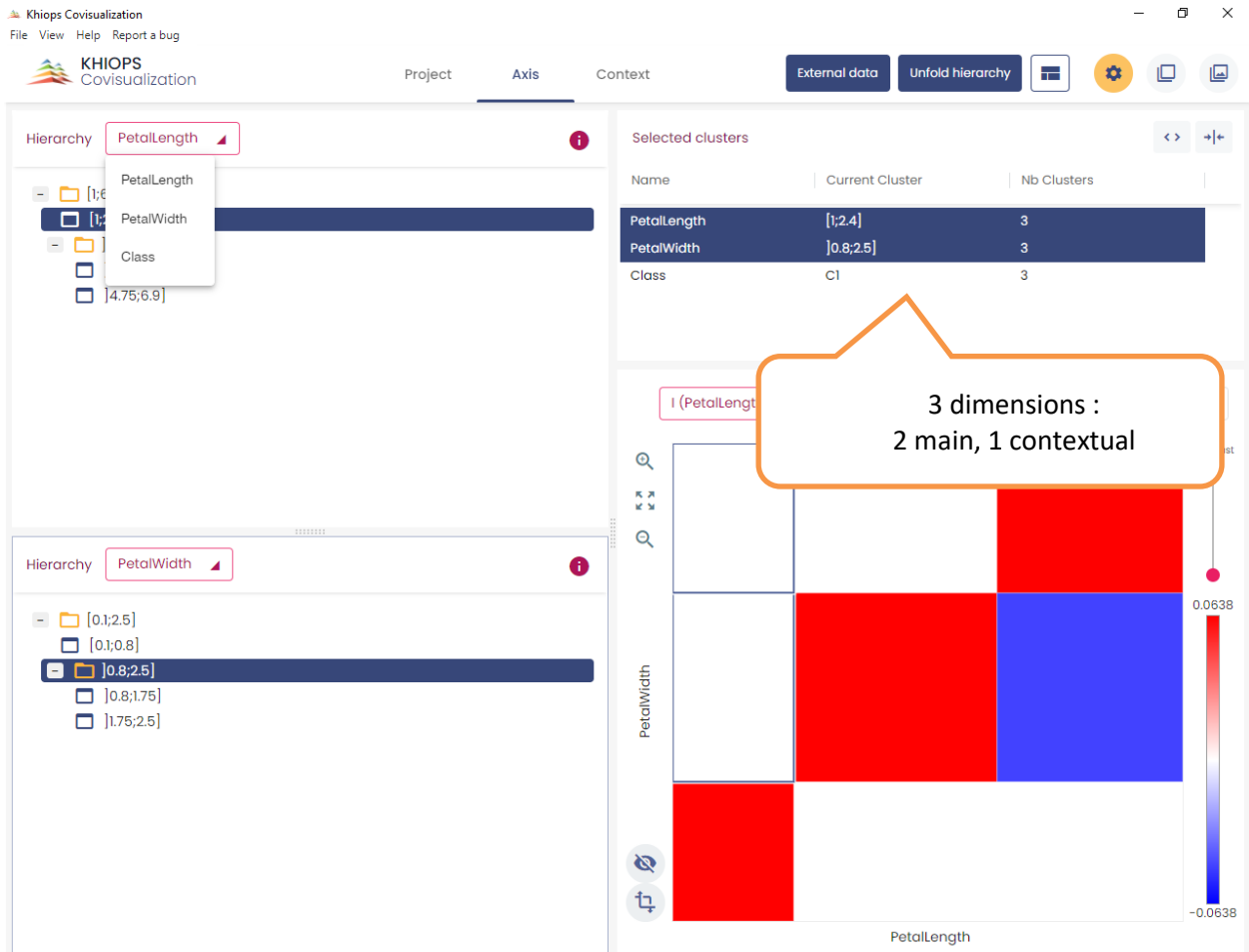
A mouse click on a cell propagates to all views



A mouse click on a cluster propagates to all views on the same dimension

5. Managing three dimensions or more

With three (or more) dimensions, the two "main" dimensions are presented as if there were only two dimensions. The other dimensions are contextual, you can view them one by one in the "context" tab.



The two dimensions of the co-occurrence matrix are always the main dimensions. By selecting a cluster in a contextual dimension, you project the co-occurrence matrix on this cluster. It allows to view two dimensions on the matrix and to act on a third. For example, with a temporal dimension as context, you can see the evolution over time of the co-occurrence matrix of two other variables.

The indicators of the cells can be computed in two ways :

- If the "Conditional on context" button is checked, the indicators are computed with the instances associated to the context selected cluster
- If it is not checked, the indicators are computed with all the instances (and have the same values whatever the context selected cluster)

Khiops Covisualization
File View Help Report a bug

KHIOPS
Covisualization

Project Axis Context

External data Unfold hierarchy

Hierarchy Class

- C1
- {Iris-setosa}
- C4
- {Iris-virginica}
- {Iris-versicolosa}

| Name | Current Cluster | Nb Clusters |
|-------------|-----------------|-------------|
| PetalLength | [1;2.4] | 3 |
| PetalWidth |]0.8;2.5] | 3 |
| Class | C1 | 3 |

I (PetalLength, PetalWidth) Conditional on context Standard Frequency

Contrast 0.0638 -0.0638

PetalWidth

PetalLength

“Class” is the contextual dimension. The axes of the matrix are the two other dimensions.

Indicators are computed according to the entire dataset or conditionally to the context

6. Technical limits

In order to be responsive on very large files, some display limitations are taken into account in Khiops Covisualization:

- The khcj file size should not exceed 400 Mo.
- The “composition” view displays the first 1 000 items. However the “copy data” feature allows to copy the first 100 000 values.
- Khiops Covisualization evaluates the available memory and displays the cocurence matrix only if there is enough available memory.
- There are slight occasional problems of automatic refreshment for large matrices. In this case you can easily refresh the matrix by switching the axis representation (Cf. 4.7.2. Axis representation).

If the khcj file is too big to be visualized in Khiops Covisualization, you can simplify the coclustering with Khiops Coclustering (see the Khiops Coclustering Guide section 3).