PROJET FOUILLE DE DONNÉES

Données socio-économiques des pays du monde en 2007

H4312

Auteurs
Stefana Gartu
Billy Pitiot

Table des matières

1 JUSTIFICATION DES CHOIX

1.1 Contexte

Nous allons étudier les données fournies par « *The World Bank Group* » en 2007 sur 209 pays du monde. Plusieurs jeux de données nous ont été fournis : différents sous-ensembles du jeu initial (countries2007_all.csv) qui ne contiennent pas de données manquantes.

Dans un premier temps, nous nous sommes intéressés aux données telles qu'elles nous ont été fournies. Nous sommes vite arrivés à la conclusion qu'il y a trop de données pour les étudier toutes dans un intervalle de temps restreint et par conséquent nous avons décidé de sélectionner un sous-ensemble de ces données. Nous avons choisi d'utiliser le fichier original avec les 209 pays et les 48 attributs, à partir duquel nous avons mené nos études en réduisant le nombre d'attributs selon ces dernières. De ce fait, nous pourrons étudier les variations entre les différentes valeurs de ces attributs et en tirer des conclusions quant aux données concernant les différents pays. Puis dans un deuxième temps, nous allons faire du clustering afin de chercher à déterminer les différents clusters qui peuvent exister dans ce jeu de donnée et d'en étudier les différentes caractéristiques. Enfin, nous effectuerons des classifications et nous étudierons à leur tour les différentes classes ainsi obtenues.

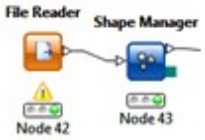
1.2 Jeux de données

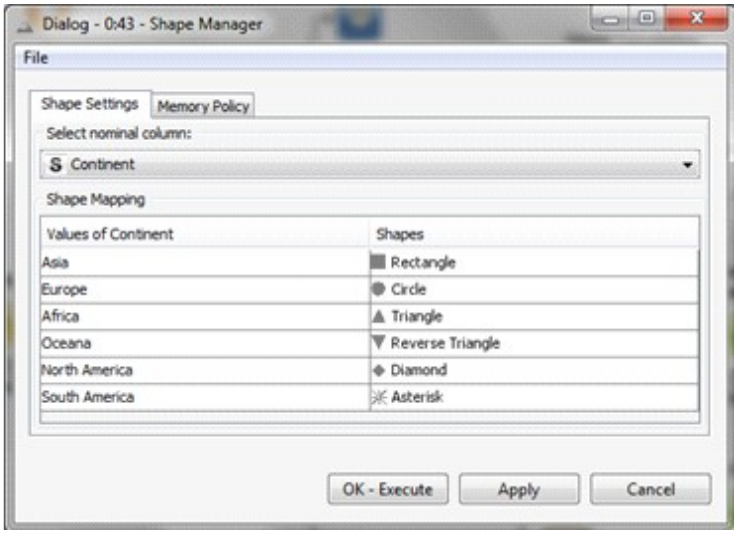
Nous avons décidé d'ajouter l'attribut *Continent* qui associe à chaque pays son continent. Les modifications ont été faites à la main directement dans le fichier source en utilisant des données trouvées sur Internet. Nous commençons donc notre traitement en définissant une forme différente (« *Shape Manager* ») pour chaque continent afin de faciliter la visualisation des résultats.


Table

utput

Country Name	Continent
Afghanistan	Asia
Albania	Country Name
Algeria	Africa
American Samoa	Oceana
Andorra	Europe
Angola	Africa
Antigua and Barbuda	North America
Argentina	South America
Armenia	Asia
Aruba	North America
Australia	Oceana
Austria	Europe
Azerbaijan	Asia
Bahamas, The	North America
Bahrain	Asia
Bangladesh	Asia

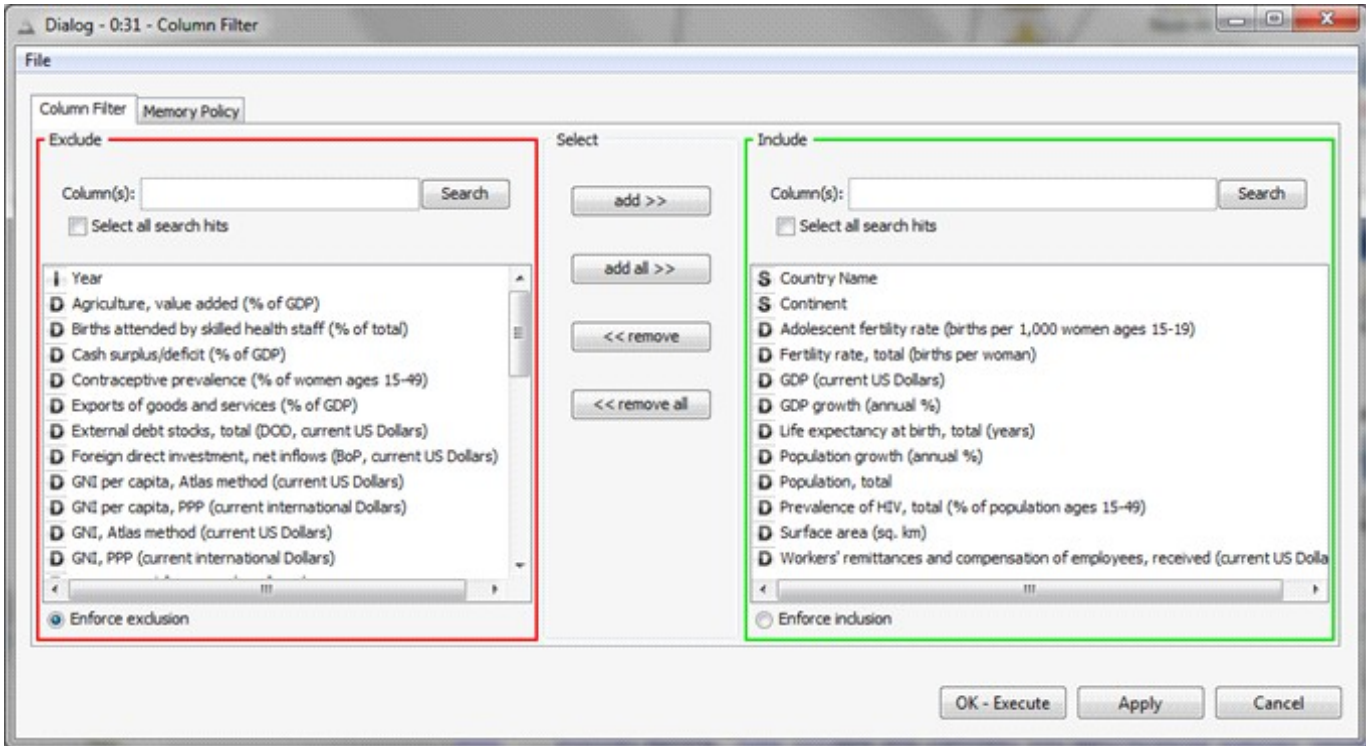




Nous avons obtenu un tableau de 209 lignes qui contient des informations sur chaque pays, parmi lesquelles le nom du continent associé. Remarquons que les pays sont inégaux dans leur description, i.e. certains pays possèdent un grand nombre de valeurs définies alors que d'autres possèdent un grand nombre de valeurs manquantes. Nous allons utiliser le filtrage de données pour remédier à ce problème.

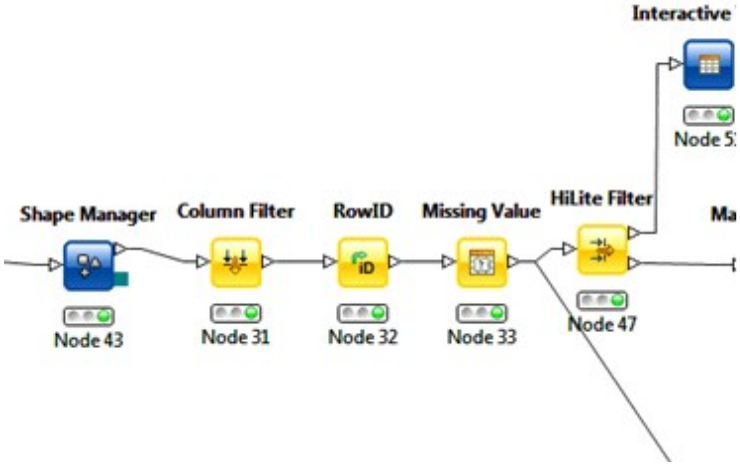
A la suite de *Shape Manager* nous allons utiliser le composant *Column Filter* où nous sélectionnions uniquement les attributs qui nous intéressent dans le cadre de notre étude. Nous commençons par la sélection de 12 attributs (mais qui vont être encore filtrés par la suite). *A noter que ces attributs seraient des critères susceptibles d'avoir une influence sur*

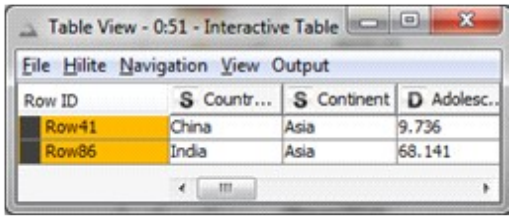
la problématique SIDA.



A la suite de ce filtrage nous avons généré une nouvelle colonne d'identification (composante *RowID*) puis filtré à nouveau les données qui contiennent des valeurs manquantes sur les attributs qui nous intéressent (composant *Missing Values*). Nous avons quand même dû faire des compromis car par exemple, un des attributs qui aurait été utile pour notre analyse (mais finalement pas retenu) est le salaire moyen dans chaque pays. Le problème est que cette donnée est renseignée pour un nombre réduit de pays et donc nous devions écarter ces pays de notre étude.

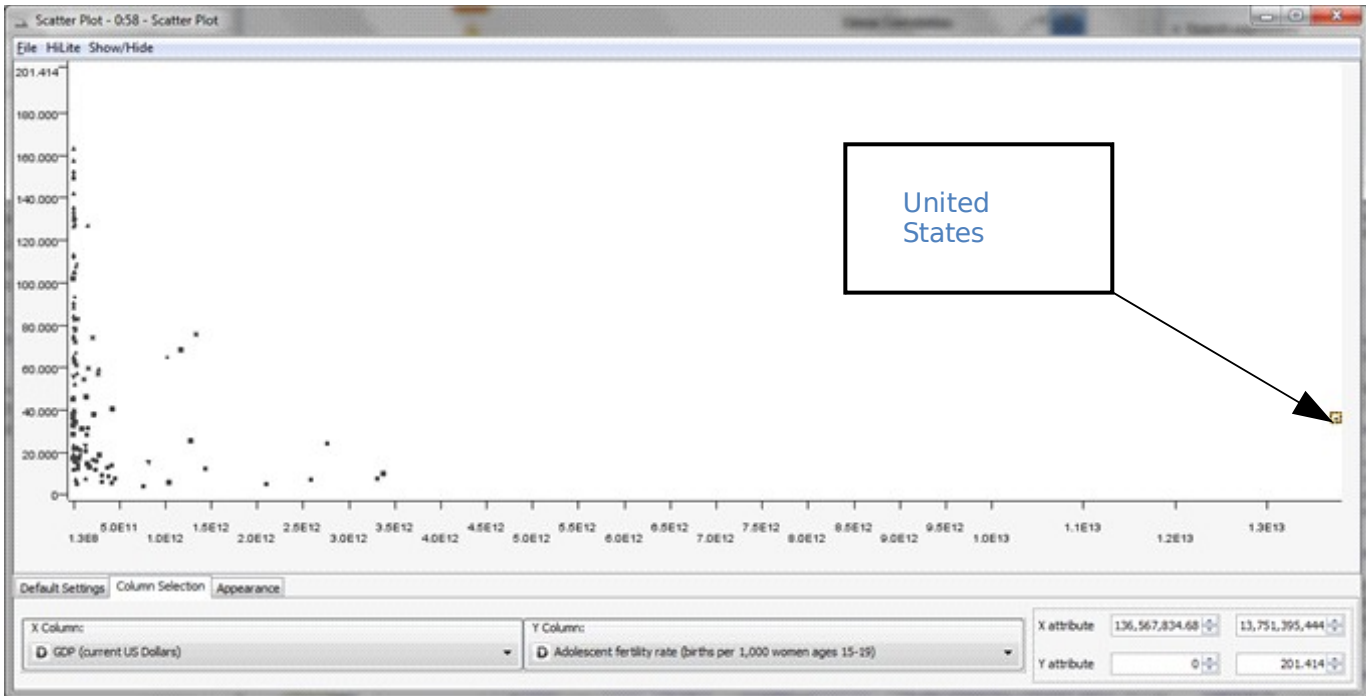
Les données telles quelles ne sont pas encore exploitables car il reste encore des extrêmes qui peuvent fausser nos résultats. En fonction des études que nous allons mener, nous allons ainsi supprimer les pays « outliers » à l'aide du composant *HiFilter* de Knime. Pour cela nous plaçons en sortie de notre source de données (sortie du composant *Missing Values)* un composant *HiLite Filter*. Celui là nous permet de séparer en deux parties notre jeu de données et donc éliminer les pays qui peuvent poser problème (par exemple la Chine dans une étude qui inclut l'attribut population).



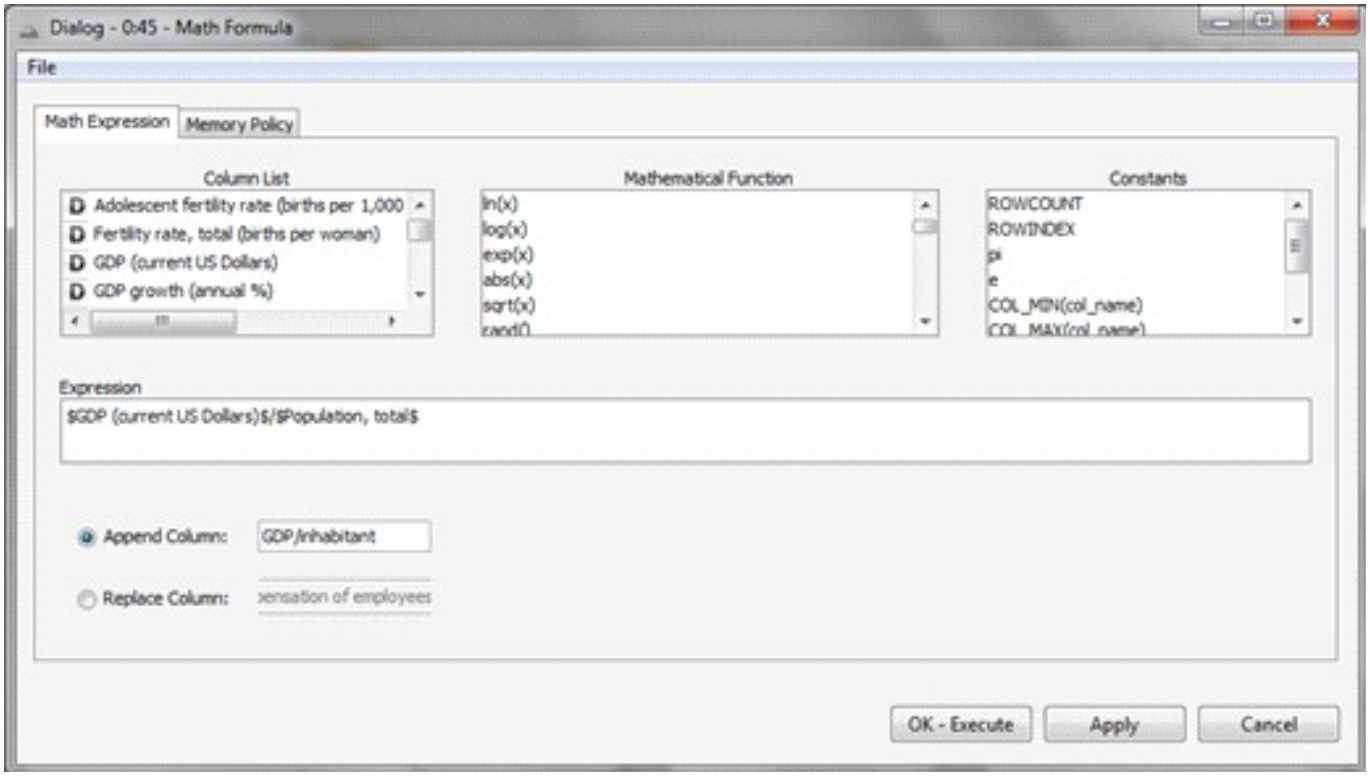


1.3 Nouvelles données

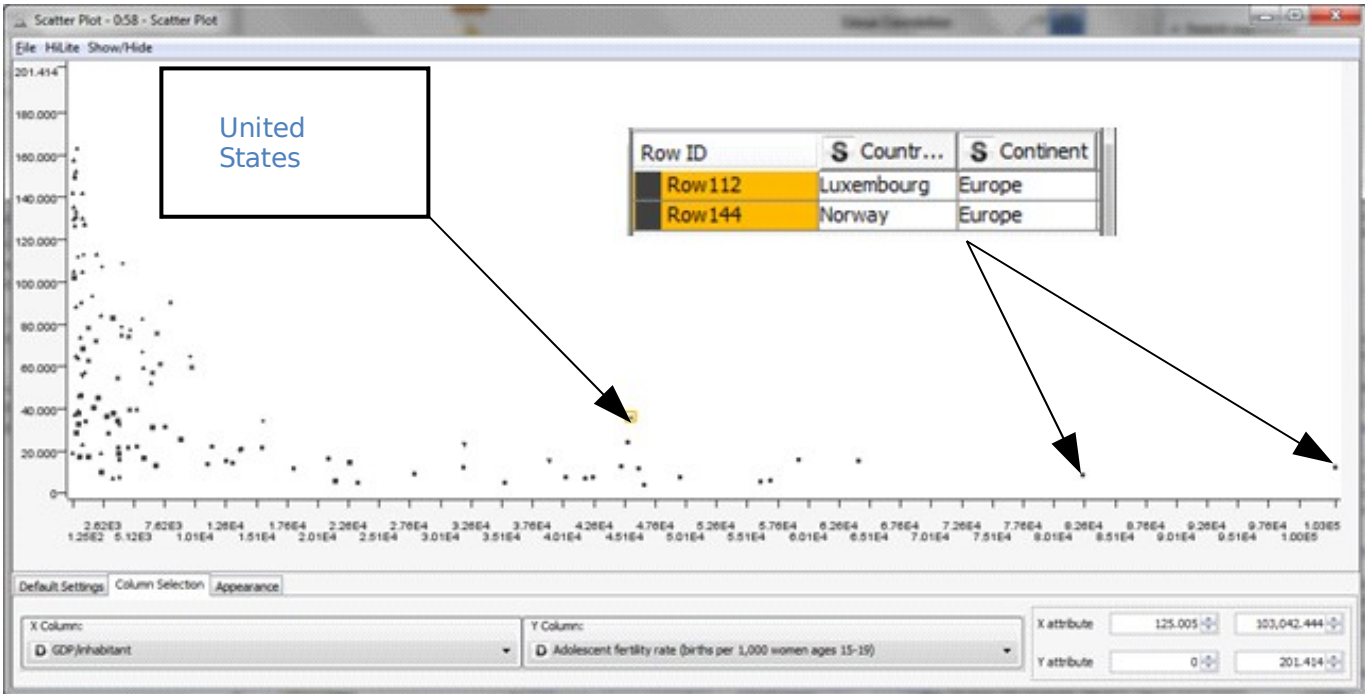
En étudiant les données obtenues nous nous sommes rendu compte que des nouveaux indicateurs utiles pourront être déduits à partir des données existantes. Par exemple le GDP des États-Unis est beaucoup plus grand que tout les autres GDP du monde. Une solution simple serait d'éliminer ce pays mais dans le même temps nous considérons que les États-Unis sont un acteur important et qu'ils devraient être étudiés avec les autres pays.



A l'aide du composant *Math Formula* nous avons calculé un nouvel indicateur **GDP/inhabitant** qui est le rapport entre le GDP et le nombre d'habitant.



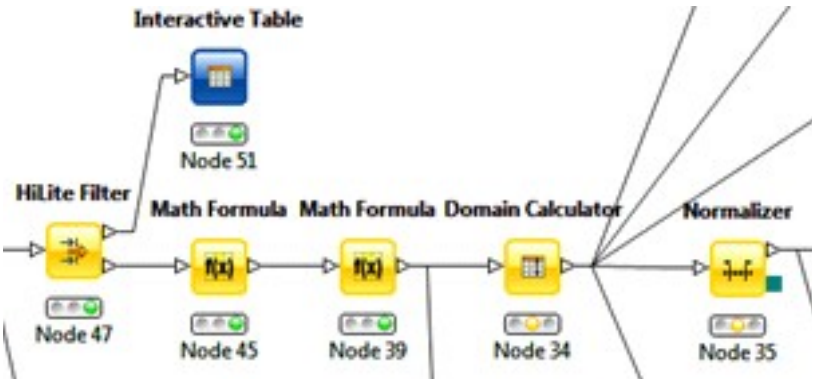
En prenant en compte ce nouvel indicateur, les Etats-Unis se retrouvent parmi les pays dits riches, mais ils ne sont plus si éloignés des autres pays du monde qu'auparavant.



Dans ce cas là, nous aurons d'autres pays qui vont s'éloigner et que nous devrons éliminer pendant nos études.

De même, nous avons créé l'indicateur **Densité** qui est spécifique à chaque pays et qui représente la population totale du pays divisée par la superficie. Dans ce cas, il y a Malte qui se distingue des autres pays avec une population de 409 000 habitants pour seulement 320 km².

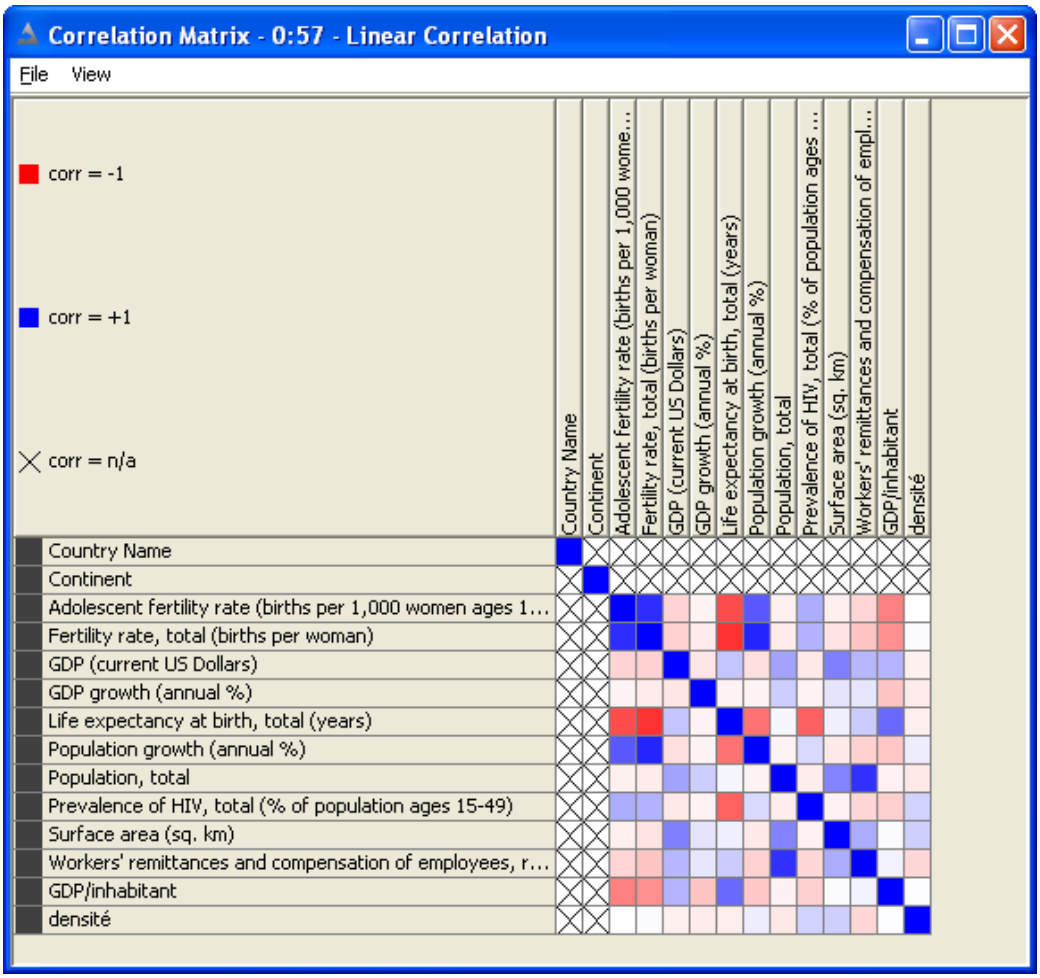
Après avoir ajouté ces nouveaux chiffres nous avons besoin d'utiliser le composant *Domain Calculator* pour bien répartir les graphiques et ne pas prendre en considération les valeurs éliminées. Ensuite il est important de normaliser les attributs du jeu de données. Cela est nécessaire avant de faire du clustering car il y a des attributs dont les domaines sont vraiment très différents. Par exemple certains attributs sont mesurés en pourcentage (de 0 à 1), tandis que d'autres attributs comme la population, s'élèvent à des valeurs de l'ordre des millions et même milliards. Pour normaliser le jeu de données nous avons utilisé le composant *Normalizer*.



2 ÉTUDE DES RÉSULTATS

2.1 Étude préliminaire

Dans un premier temps, nous avons cherché à voir si nos attributs n'étaient pas redondant. Nous avons pour cela étudié la corrélation linéaire entre les différents attributs.



Les attributs ne sont pas trop corrélés car nous avions d'office éliminé ceux qui nous semblait redondant et donc non intéressant tels que les différents GNI/GDP. Nous n'avons gardé que la fertilité et la fertilité limitée aux jeunes femmes qui sont très corrélées car elles sont liées à ce qu'on voulait étudier à la base : le lien entre le pourcentage de personnes atteinte par le SIDA et les autres attributs. Il y a aussi la croissance de la population qui est très corrélée à la fertilité, ce qui semble logique et donc par la suite, nous ne l'avons pas pris en compte pour la création des clusters ou pour leur interprétation.

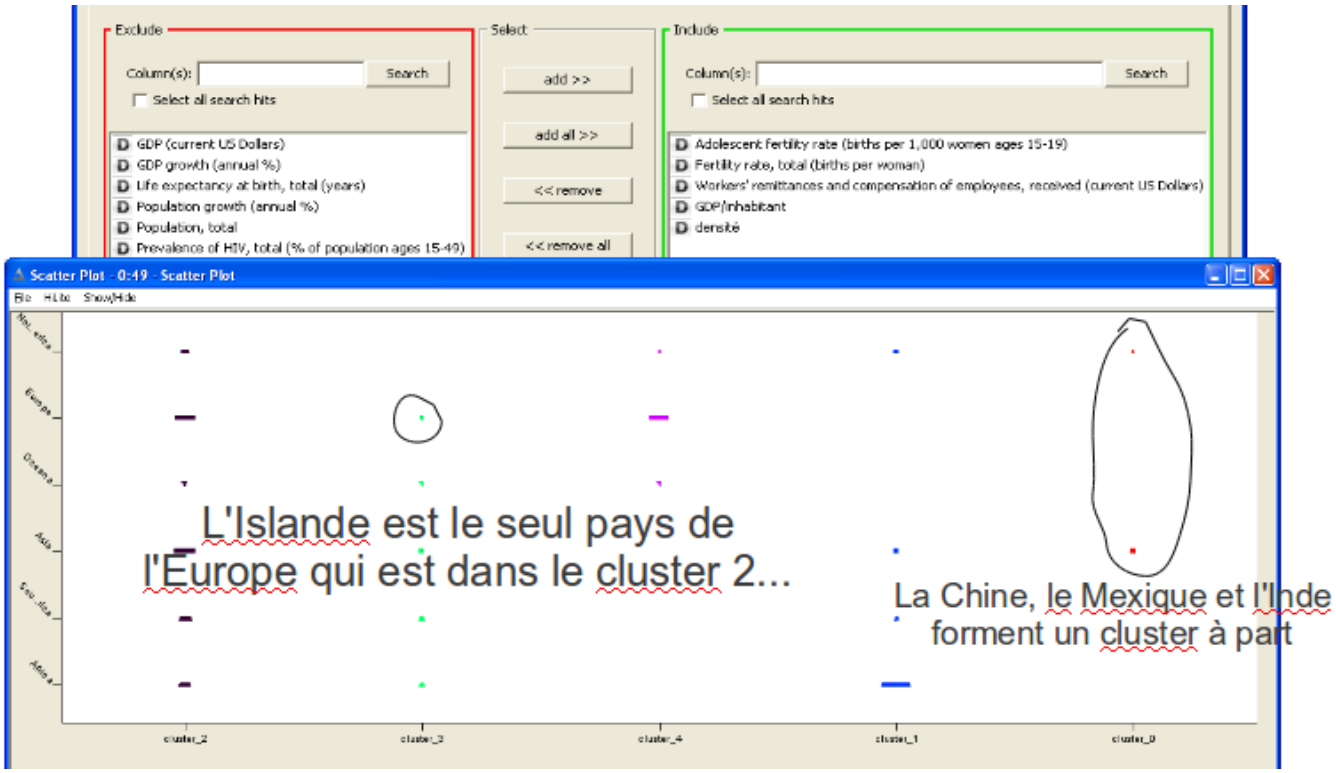
2.2 Clustering

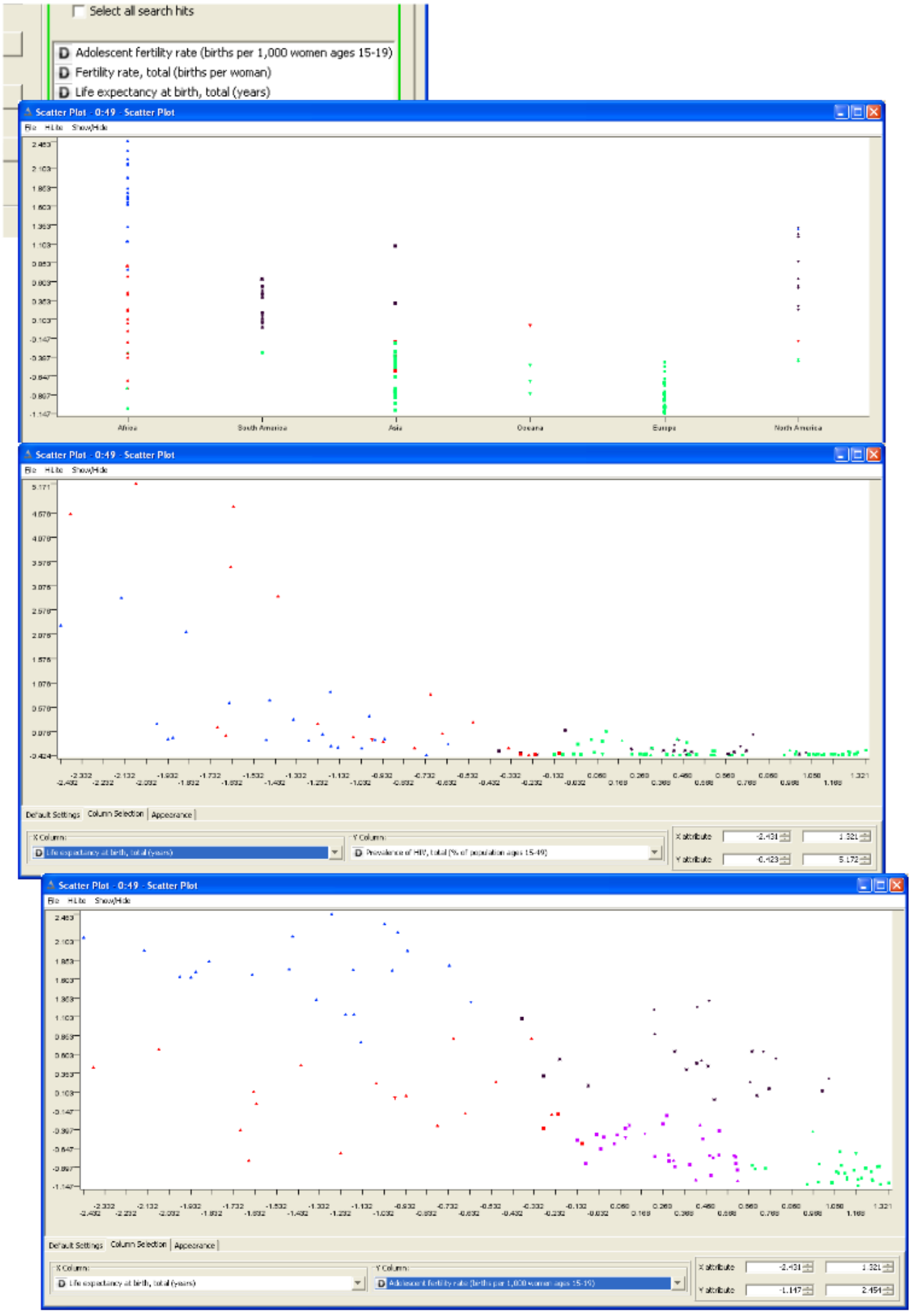
Pour réaliser du clustering, nous avons principalement utilisé les composants *hierarchical cluster* et *k-means*. Par manque de temps, nous n'avons pas pu utiliser d'autres méthodes de clustering mais il aurait pu être intéressant de comparer les résultats obtenus grâce aux autres méthodes.

La méthode du *hierarchical clustering* ne s'est pas révélée très intéressante dans notre cas car elle nous donnait systématiquement un cluster de quelques pays à coté d'un cluster énorme contenant tous les autres. Même en supprimant les *outliers*, avec les différents attributs qui nous intéressaient et en utilisant les différentes méthodes (distance min, max et moyenne), cette répartition était à chaque fois retrouvée et nous n'avons pas

réussi à l'interpréter .

Au contraire, la méthode de *k-means*, s'est révélée intéressante et nous a fourni des résultats que nous avons pu essayer d'interpréter.





Au départ, nous ne savions pas trop comment chercher, ...

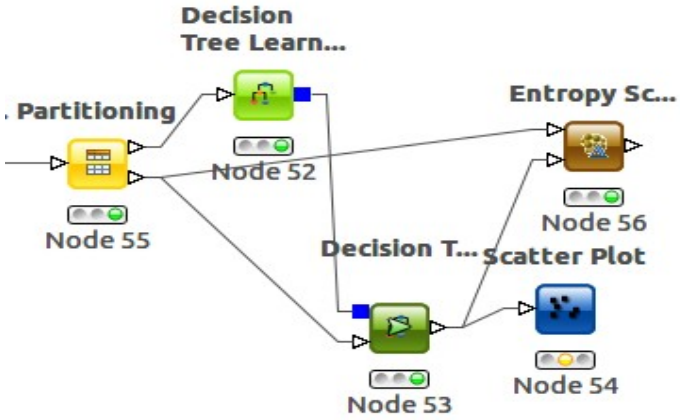
Dans les graphiques ci-dessus, nous pouvons observer plusieurs choses intéressantes. Par exemple, les trois pays en verts appartenant à l'Afrique sont les trois pays du Maghreb : le Maroc, la Tunisie et l'Algérie. En cela, on voit qu'ils sont bien plus proche de l'Europe avec qui ils ont de nombreux lien grâce à la méditerranée que des autres pays d'Afrique.

Il aurait pu être intéressant de tester si k-means était résistant à plusieurs exécutions avec le composant loop plutôt qu'avec nos tests répétés mais à cause d'un manque de temps, nous nous sommes limités à effectuer plusieurs tests et à regarder les résultats à chaque fois.

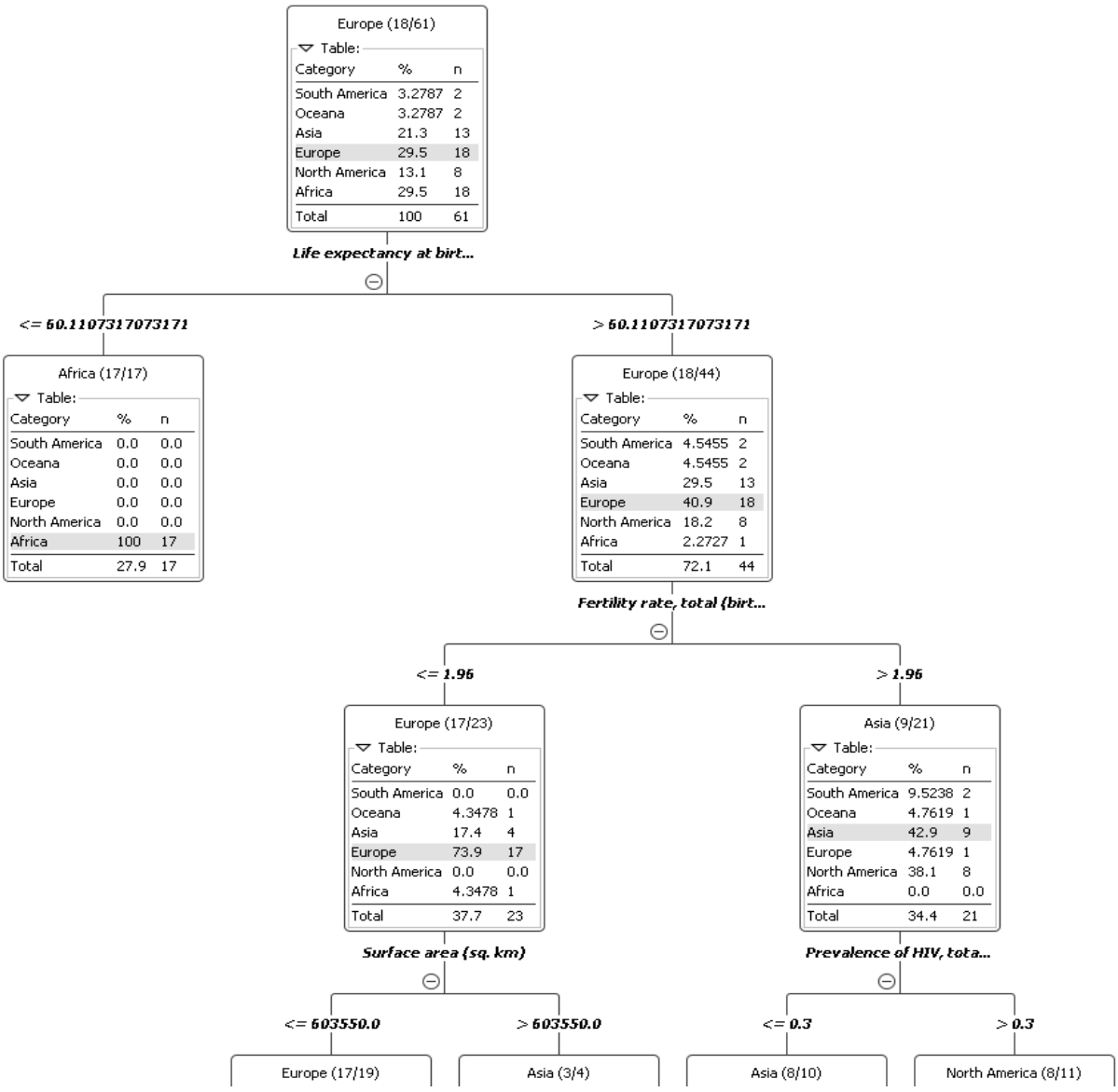
2.3 Classification

L'interprétation étant quelque chose de compliqué, souvent subjectif et nous ayant fait perdre beaucoup de temps sans que l'on ai l'impression d'être efficace, nous avons décidé d'essayer de faire de la classification. En fait, nous l'avons surtout décidé parce qu'on s'est aperçu au cours de nos recherches que nos clusters correspondaient souvent assez bien aux continents à quelques exceptions près. Notre but était donc de voir quels attributs permettait de retrouver les continents à l'aide d'une classification.

Nous avons donc essayé de mettre un *decision tree learner* et de regarder l'arbre de décision ainsi obtenu après être passé dans un composant *partitioning* envoyant 90% des données choisies aléatoirement dans l'arbre de décision.



Notre schéma comprend donc le composant *Partitioning* qui permet de répartir les données aléatoirement, les *Decision Tree Leaner* et *Predictor* qui permettent de construire l'arbre de prédiction



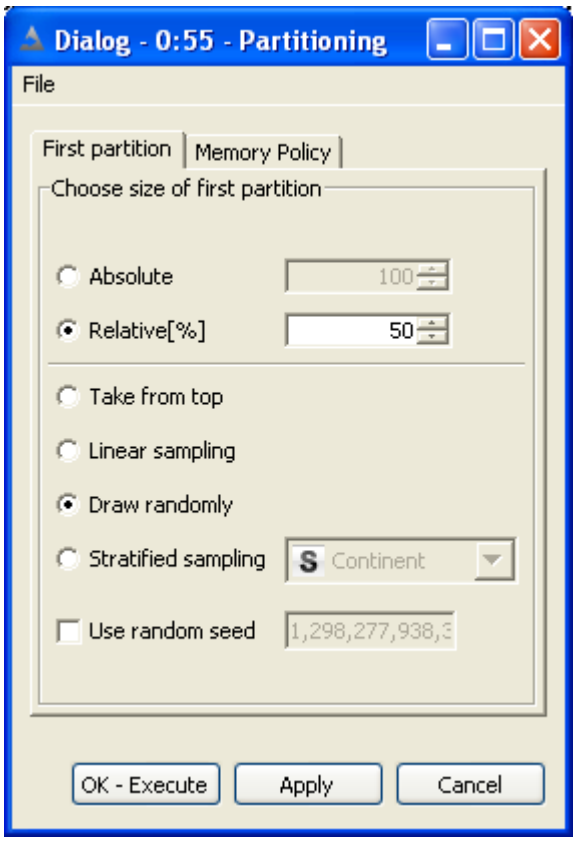
Nous pouvons observer qu'avec l'espérance de vie à la naissance, nous pouvons directement mettre de coté 17 des 18 pays d'Afrique qui ont été testé. Ceci est légèrement contradictoire avec ce que nous avions vu précédemment grâce au clustering : les pays du maghreb sont à part. Cette contradiction vient du fait que les tests ne sont pas effectués sur l'ensemble des données et apparemment ici, seul un des trois pays apparaît. C'est tout de même très intéressant de voir apparaître une telle différence entre l'Afrique et les autres pays.

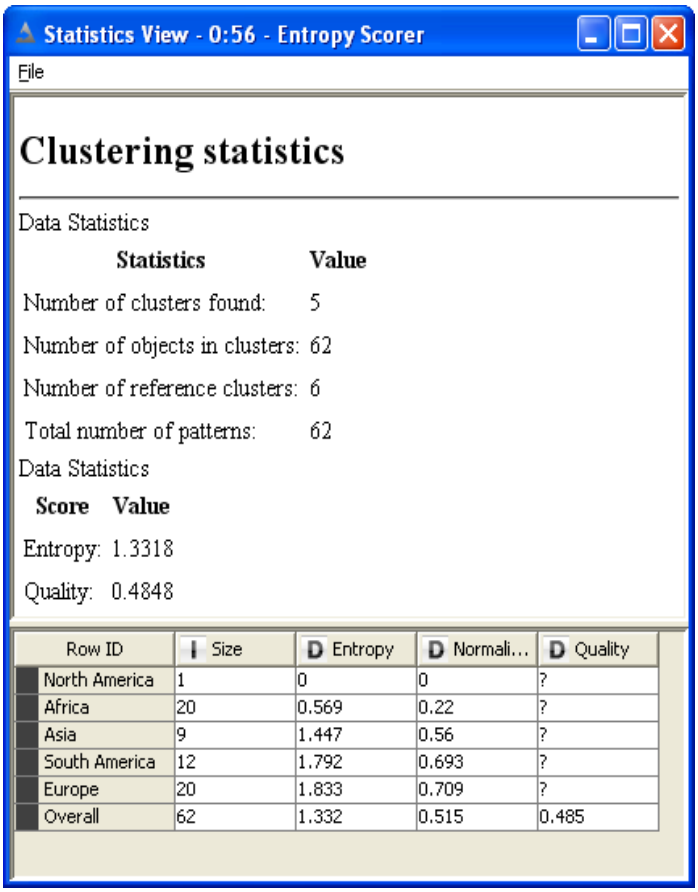
Ensuite, nous voyons apparaître presque tous les pays d'Europe lorsque l'on prend le nombre de naissance par femme. Avec ces pays, nous retrouvons le dernier pays d'Afrique, ce qui vient corroborer nos découvertes précédentes et quelques pays d'Asie. Nous ne sommes pas rentré beaucoup plus dans le détail de ce coté, nous pouvons juste voir que la répartition suivante se fait en fonction de la superficie des pays avec d'un coté les pays asiatiques et de l'autre l'Europe.

Par contre, on peut se pencher sur la branche de l'arbre restante. Dans cette branche, nous voyons que pour départager les 21 pays restants, il faut regarder le nombre de personnes atteinte par le SIDA. Ce nombre est inférieur à 30% pour les pays asiatiques et supérieur pour les pays d'Amérique.

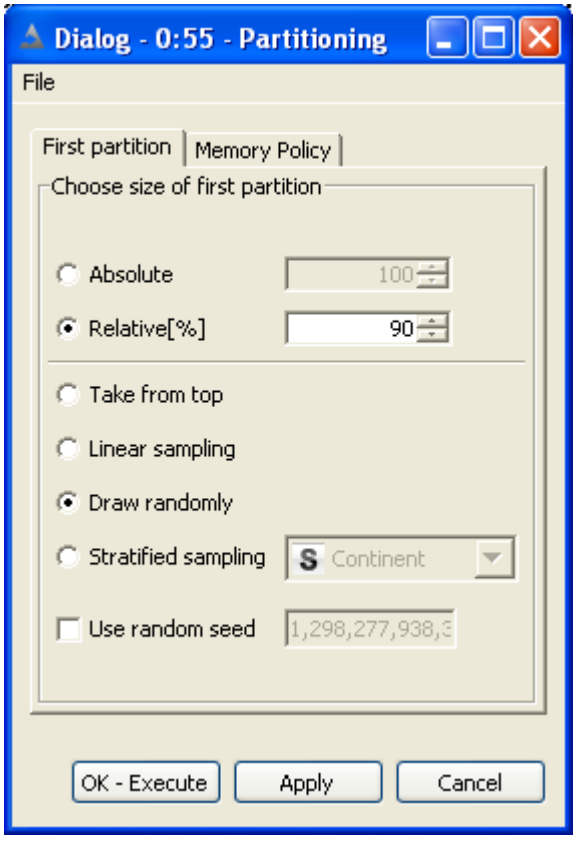
Toutes ces informations sont vraiment intéressantes dans le sens où ce ne sont pas des données acquise ou évidentes. Et elles permettent à quelques exceptions prêtes de séparer clairement les différents continents.

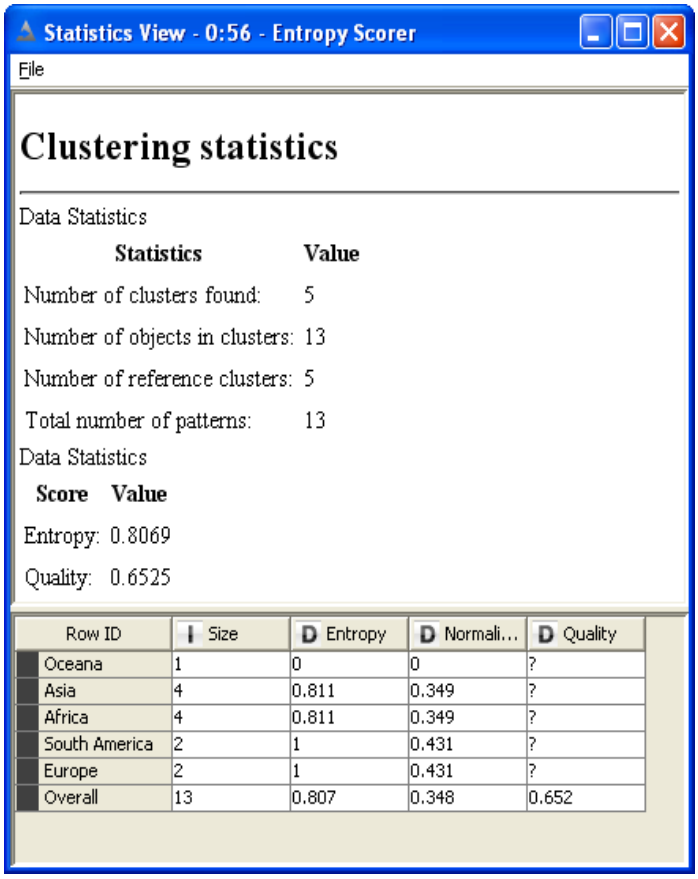
Par la suite, nous avons mis en œuvre ce *Predictor* et nous lui avons demandé de prédire les continents d'autres pays. Nous obtenons les résultats suivants en sortie de *l'Entropy Scorer* :





Row ID	Size	Entropy	Normali...	Quality
North America	1	0	0	?
Africa	20	0.569	0.22	?
Asia	9	1.447	0.56	?
South America	12	1.792	0.693	?
Europe	20	1.833	0.709	?
Overall	62	1.332	0.515	0.485





Row ID	Size	Entropy	Normali...	Quality
Oceana	1	0	0	?
Asia	4	0.811	0.349	?
Africa	4	0.811	0.349	?
South America	2	1	0.431	?
Europe	2	1	0.431	?
Overall	13	0.807	0.348	0.652

Nous pouvons remarquer que l'entropie est plutôt bonne lorsqu'on construit l'arbre avec 90% des données mais lorsqu'on le construit avec 50% l'entropie chute de moitié et les résultats ne peuvent plus être acceptés avec confiance.