El AMRANI Rachida **2005/2006**
RAKOTOMANGA Harijaona

PROJET SODAS
Analyse de données sur les livres

Université de Paris Dauphine **Master Info T.I.O.**
Data Mining

Sommaire

Introduction

Face à une concurrence accrue dans tous les secteurs, une entreprise doit pouvoir anticiper face à ces concurrents. Pour cela, une entreprise doit disposer d'informations pertinentes.

En effet, aujourd'hui, nous sommes capables de stocker un nombre très impressionnant de données quelque soit le format de ces derniers. Ainsi, bon nombre de ces données peuvent et sont souvent stockées à des fins d'analyse.

Afin de pour voir en extraire des informations pertinentes il est nécessaire de les regrouper de façon subtile. A partir d'une quantité impressionnante de données, d'informations, le datamining permet de dégager des informations qui ne sont pas visibles à l'état brut.

Pour cela SODAS est un logiciel qui permet à l'aide de concept d'extraire des informations pertinentes et d'avoir une connaissance plus approfondie, en recherchant des relations et de structurer les données.

I. Contexte : Datamining - SODAS

A. Le Datamining

Etant donné le nombre impressionnant de données qu'une entreprise, ou autre, peut stocker, il est nécessaire de pouvoir les regrouper, en concept, afin de pouvoir en extraire le maximum d'informations. Tout ceci dans le but d'analyser, synthétiser et organiser les données ainsi que les informations. Ces concepts peuvent être modélisés par des données de niveaux supérieurs, par des données symboliques.
Ces derniers ont pour but de décrire les concepts et également, de les décrire au sein de l'ensemble des individus qu'ils représentent.

Ces informations peuvent être divisées en deux catégories ou deux niveaux : les informations relatives à une ligne, un individu d'un ensemble X, puis les concepts, un niveau agrégé.

B. SODAS

Est un logiciel qui permet de traiter des données a l'aide de données symboliques. Il permet, à partir d'une base de données de construire un tableau de données symboliques, grâce à des règles et des taxonomies, dans le but de décrire des concepts.

Chaque concept est décrit à l'aide de variables (de différentes natures). Ainsi, grâce au logiciel SODAS, un fichier d'objets symboliques auquel on peut appliquer une douzaine de méthodes d'analyses.

II. Application : Etude de données sur les livres

A. Présentation des données

1. Source des données : La base BDLivres.mdb

La base de données BDLivres.mdb est la base de données sur laquelle repose notre étude.
Il s'agit d'une base de données relationnelles sous ACCESS. Elle a été fournit par des données se trouvant sur Internet et notamment des sites référençant la liste des livres ayant obtenues des prix littéraires : Goncourt, Femina etc…

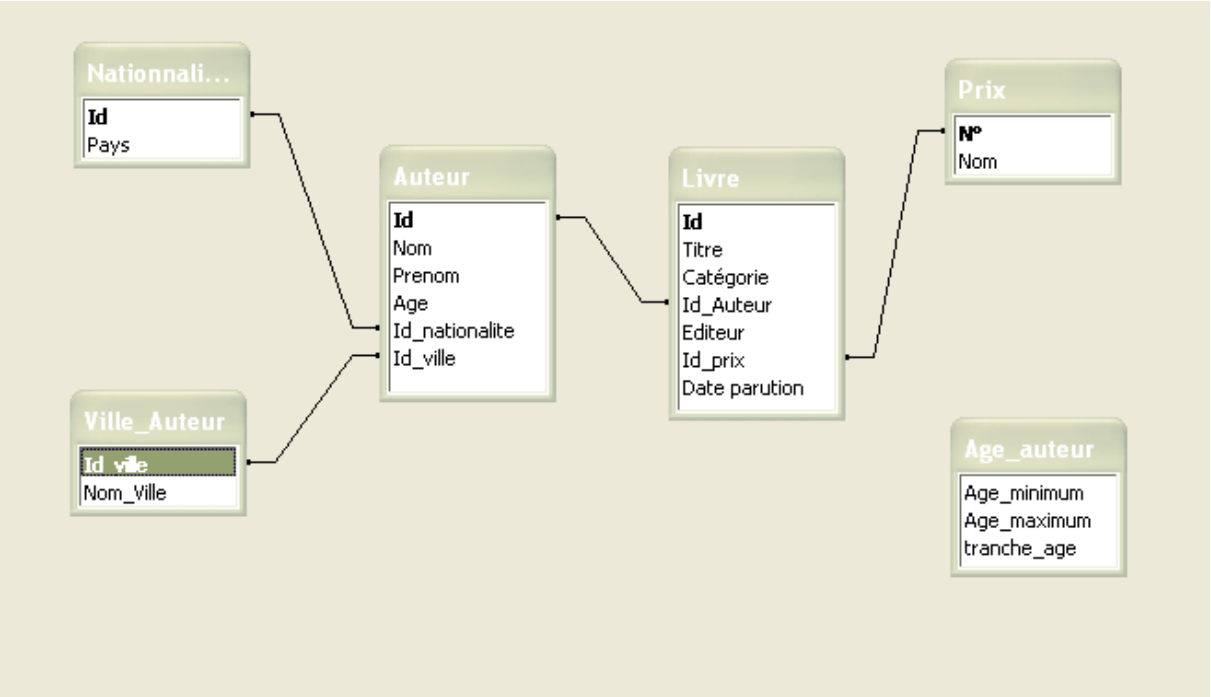
Notre base de données contient une centaine de livres qui ont été présenté à des prix littéraires. Sur chaque livre, nous avons le nom et prénom de l'auteur, la maison d'édition,le titre du livre, l'age de l'auteur etc…
Nous allons dans le prochain paragraphe présenter le schéma de la base, les différentes tables, et relations.

2. Présentation de la base

Notre base de données BDLivres.mdb est composée des tables suivantes :

- La table LIVRE contient les informations relatives aux livres :
  - Titre
  - Catégorie
  - Editeur
  - Date de parution…
- La table PRIX contient les différents prix gagnés :
  - Goncourt
  - Fémina
  - Médicis
  - Renaudot
- La table AUTEUR contient les informations relatives à l'auteur :
  - Nom
  - Prénom
  - Age
- La table AGE_AUTEUR contient les intervalles d'âge : 7 tranches d'âge de 20 à 90 ans découpé en décennies.
- La table NATIONALITE_AUTEUR contient les différentes nationalités d'origine des auteurs :
  - Française
  - Espagnol
  - Anglaise
  - Allemande
- La table VILLE_AUTEUR contient les villes d'origine des différents auteurs :
  - Paris / Marseille / Lyon
  - Berlin / Francfort
  - Madrid / Barcelone
  - Londres / Manchester

Dans notre cas, Un auteur ayant une certaine nationalité « d'origine », ne peut être originaire que d'une ville du pays de sa nationalité : exple un auteur allemand ne peut être originaire que des villes de Berlin ou Francfort.


Nationnali...
Id
Pays
Auteur
Id
Nom
Prenom
Age
Id_nationalite
Id_ville
Livre
Id
Titre
Catégorie
Id_Auteur
Editeur
Id_prix
Date parution
Prix
N°
Nom
Ville_Auteur
Id_ville
Nom_Ville
Age_auteur
Age_minimum
Age_maximum
tranche_age

B. Variables et requêtes

1. Individus, variables et concepts

Les individus sont les livres sélectionnés pour participer aux prix littéraires, notre base de données regroupe 100 individus.

Les variables de description sont :

- La date de parution
- l'age de l'auteur
- le nom de l'auteur
- la nationalité de l'auteur
- le nombre de prix pour chaque catégorie de livre
- L'éditeur du livre

Les concepts sont les différentes catégories des livres :

- Littérature française
- Littérature anglaise
- Littérature allemande
- Littérature espagnole

2. Requêtes

Nous utiliserons trois requêtes dans notre analyse :

La requête REQ_LIVRE nous renvoie la majorité des variables de descriptions, les informations relatives aux livres.

- *REQ_LIVRE*

SELECT DISTINCTROW Livre.Titre, Livre.Catégorie, Auteur.Nom, Nationnalite_auteur.Pays, Prix.Nom, Auteur.Prenom, Auteur.Age, Age_auteur.tranche_age
FROM Age_auteur, Prix INNER JOIN (Nationnalite_auteur INNER JOIN (Auteur INNER JOIN Livre ON Auteur.Id = Livre.Id_Auteur) ON Nationnalite_auteur.Id = Auteur.Id_nationalite) ON Prix.N° = Livre.Id_prix
WHERE (((Auteur.Age) Between [Age_auteur].[Age_minimum] And [age_Auteur].[Age_maximum]) AND ((Auteur.Id)=[Livre].[Id_Auteur]) AND ((Nationnalite_auteur.Id)=[Auteur].[Id_nationalite]) AND ((Prix.N°)=[Livre].[Id_prix]));

La requête REQ_CATEGORIE nous renvoie le nombre de livres ayant reçu un prix par catégorie.

- *REQ_CATEGORIE*

SELECT Livre.Catégorie, Count(Livre.Id_prix) AS Nb_prix
FROM Livre
GROUP BY Livre.Catégorie;

La requête REQ_EDITEUR insère une variable multimodale Edition (ou éditeur).

- *REQ_EDITEUR*

SELECT DISTINCT Livre.Catégorie, Livre.Editeur, 1 AS pondération
FROM Livre;

3. Taxonomies

Pour la taxonomie, la requête Taxonomie_Auteur reprend les deux variables nationalités et ville, qui sont dans notre étude soumise à une hiérarchie (cf. présentation des tables), dans la mesure ou si un auteur est de nationalité française, donc de France, il ne peut originaire que d'une ville de France, en l'occurrence au sein de notre Paris, ou Marseille ou Lyon.

- *Taxonomie TAXONOMIE_AUTEUR :*

```
SELECT DISTINCT Nationnalite_auteur.Pays, Ville_Auteur.Nom_Ville
FROM Nationnalite_auteur INNER JOIN (Ville_Auteur INNER JOIN Auteur ON Ville_Auteur.Id_ville = Auteur.Id_ville) ON Nationnalite_auteur.Id = Auteur.Id_nationalite;
```

C. Présentation et application des différentes méthodes utilisées

1. Méthode SOE

- Présentation

La méthode SOE pour Symbolic Object Editor, permet de visualiser tous les objets symboliques présent dans le fichier sodas. On peut, si on le souhaite faire des petites modifications sur les données.
SOE permet également, de visualiser chaque objet symbolique à l'aide de la représentation SOL, Symbolic Object Language, ainsi que des représentations graphiques en 2D ou 3D.

- Application

Les objets que nous analyserons sont :

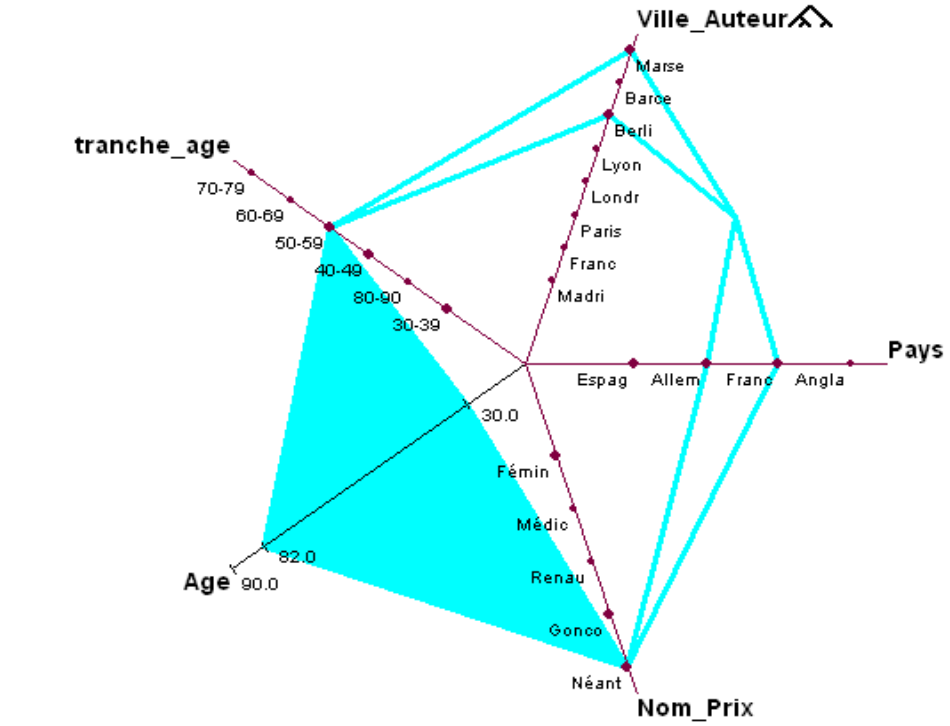
- Littérature espagnole
- Littérature française
- Littérature anglaise
- Littérature allemande

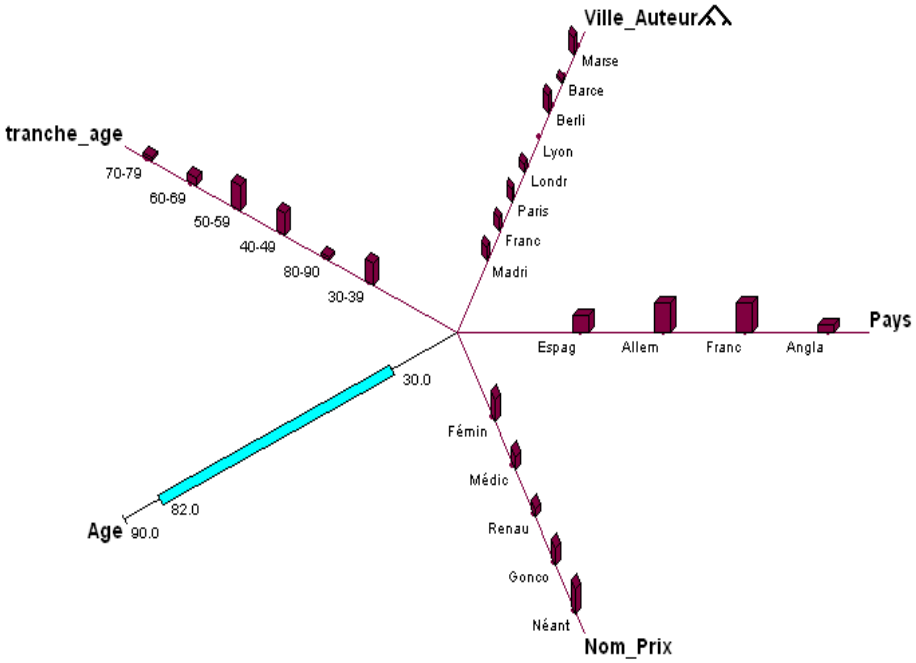
Ce sont les objets symboliques que nous allons analyser. Ces derniers sont en fait les catégories de livres de notre base.

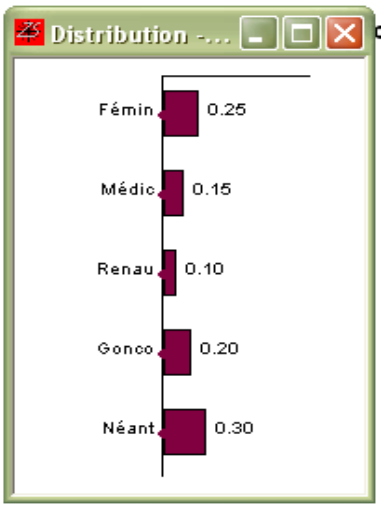
Les variables descriptives que nous avons choisis pour l'analyse de chaque objet sont :

- Age
- Tranche d'âge
- Nom du prix
- Pays de l'auteur
- Ville de l'auteur

Littérature espagnole



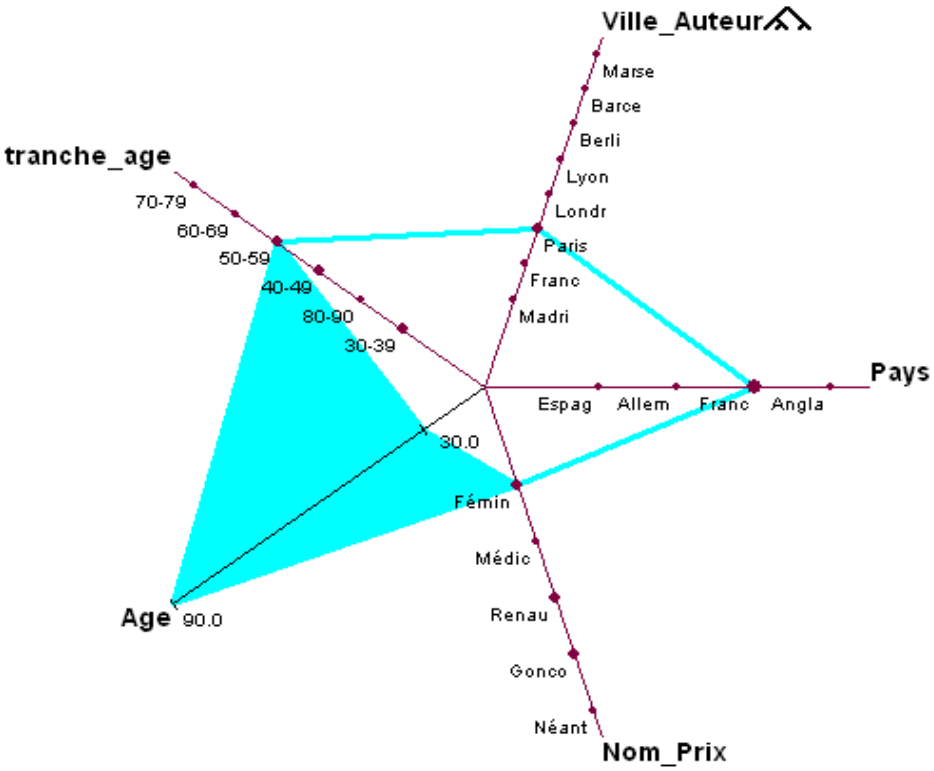


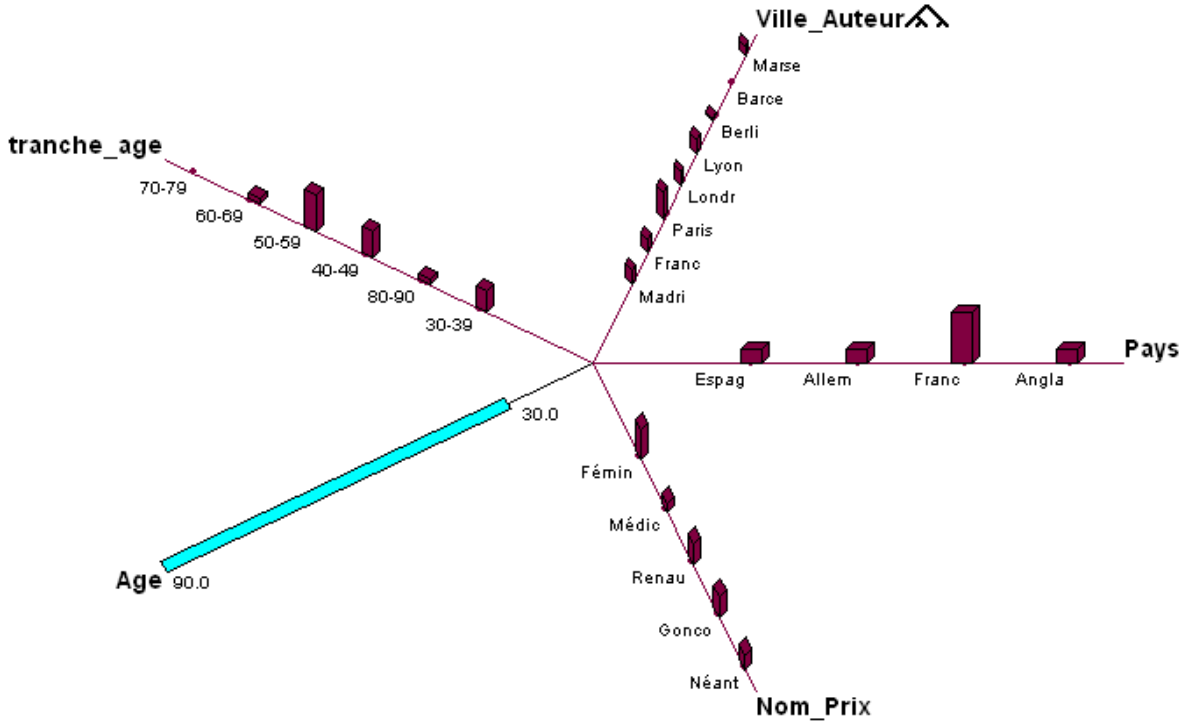


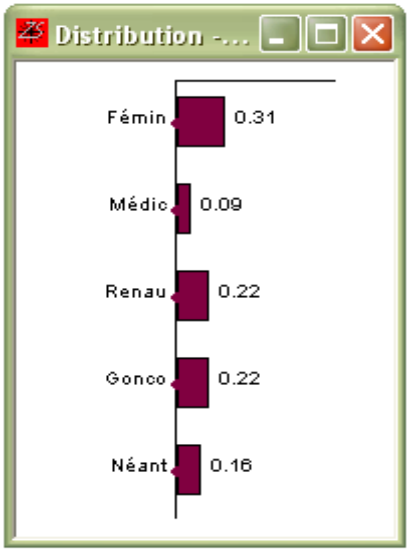
A l'aide de ces graphiques, on remarque qu'un tiers des livres de la littérature espagnole ne remportent aucun prix. De plus, les livres remportant un prix se répartissent entre le prix Fémina (25%) et le prix Goncourt (20%).

Les auteurs des livres de la littérature espagnole sont pour la plupart originaire de France et d'Allemagne, et concernant la ville, ils sont originaires pour la plupart de Marseille et de Berlin.

Littérature française



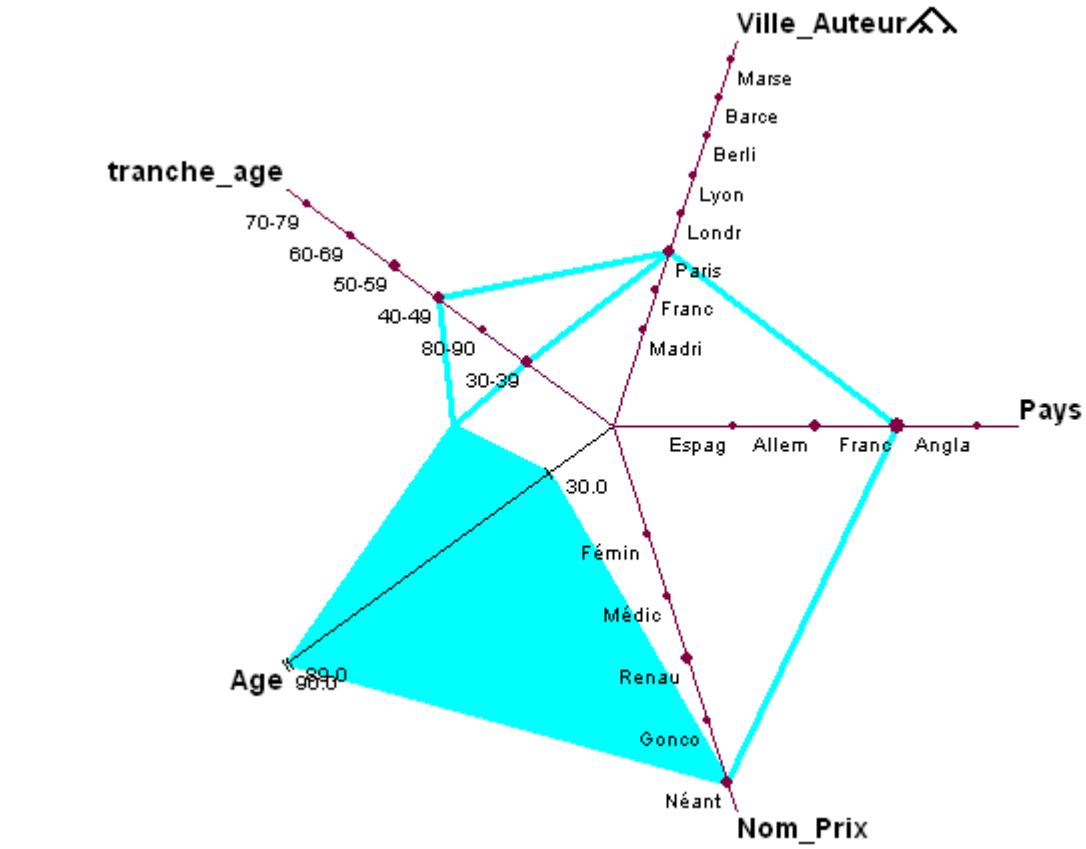


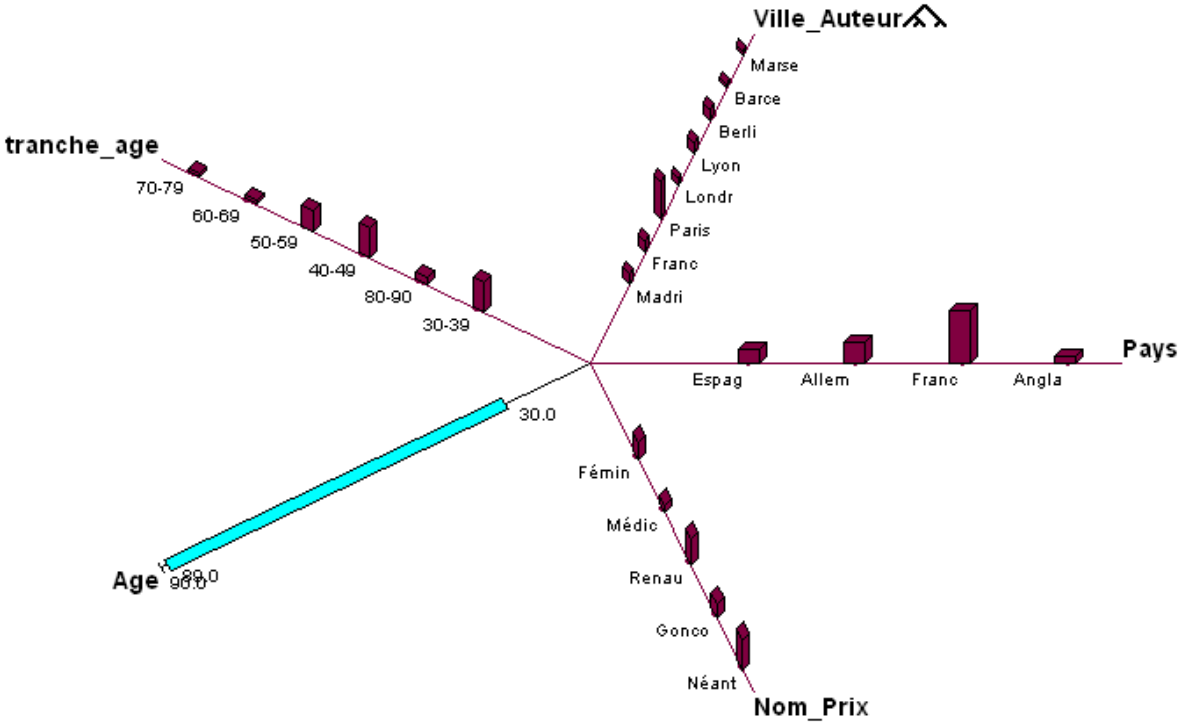


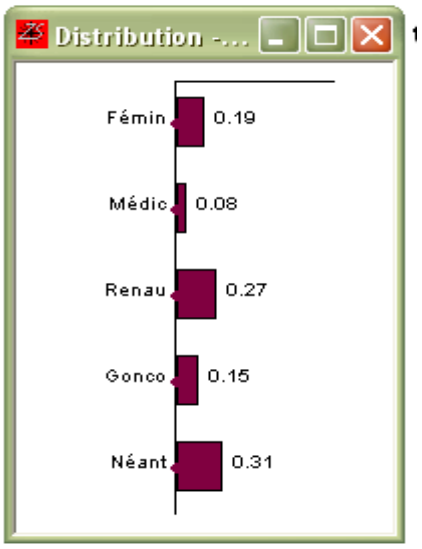
On remarque, que les auteurs de livres de la catégorie « littérature française » sont d'origine française et sont parisiens d'origine.

On remarque, également, que les livres de la littérature française remportent pour 31% le prix Fémina. Sur les livres sélectionnés en catégorie « littérature française », 84% remportent un prix contre seulement 16% qui ne remporte aucun prix.

Littérature anglaise





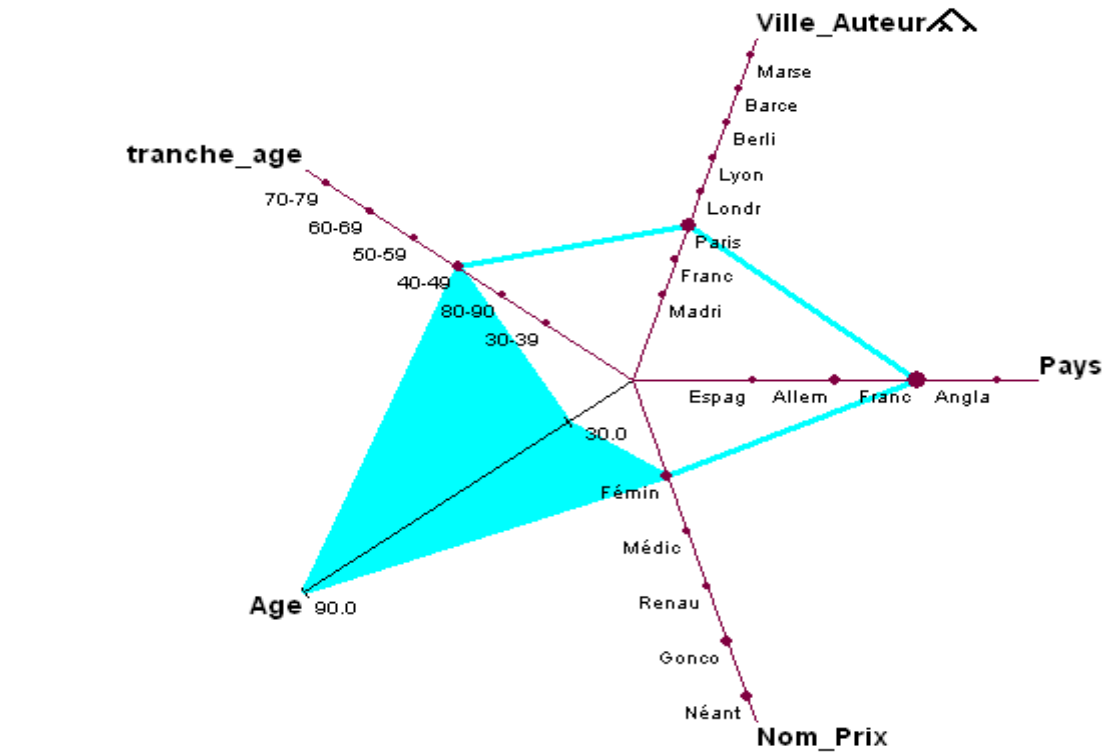


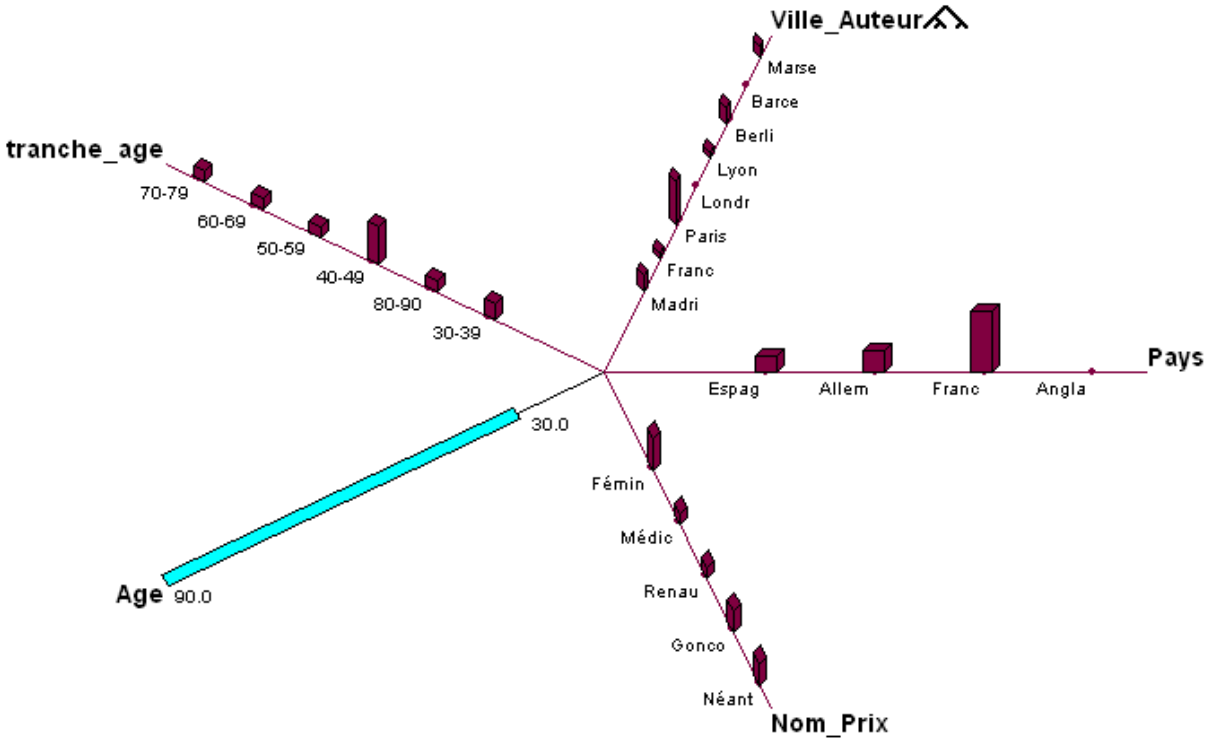
On remarque qu'un tiers des livres de la littérature anglaise ne remporte aucun prix. Pour les livres remportant un prix 27% des livres sélectionnés remportent le prix Renaudot.

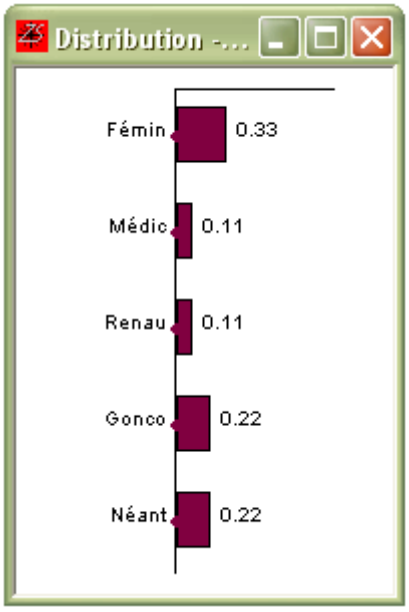
Nous remarquons également, que les livres de la catégorie littérature anglaise, sélectionnés au concourt de littérature, ont pour auteurs des personnes d'origine française, et parisienne plus précisément.

Nous remarquons que les livre sde cette catégories ont des auteurs se répartissant sur deux tranches d'âge les « 40-49 » et les « 30-39 » ans.

Littérature allemande







Interprétation générale :

Nous remarquons, toute catégorie confondue, que les livres sélectionnés au concourt de littérature sont pour la grande majorité des livres dont l'auteur est d'origine française et de la ville de Paris.

De plus nous remarquons que les auteurs sont compris entre 30 ans et 59 ans, bien qu'une grande proportion de ecs auteurs se trouvent dans la tranche d'âge 50-59 ans.

Nous pouvons constater que le prix le plus remporter est le prix Fémina, bien que certaines catégories ont le pourcentage plus élevé sur aucun prix.

Puis, toute catégorie confondue, la plus grande majorité des livres de chaque catégorie de littérature remporte un prix.

2. Méthode STAT

- Présentation

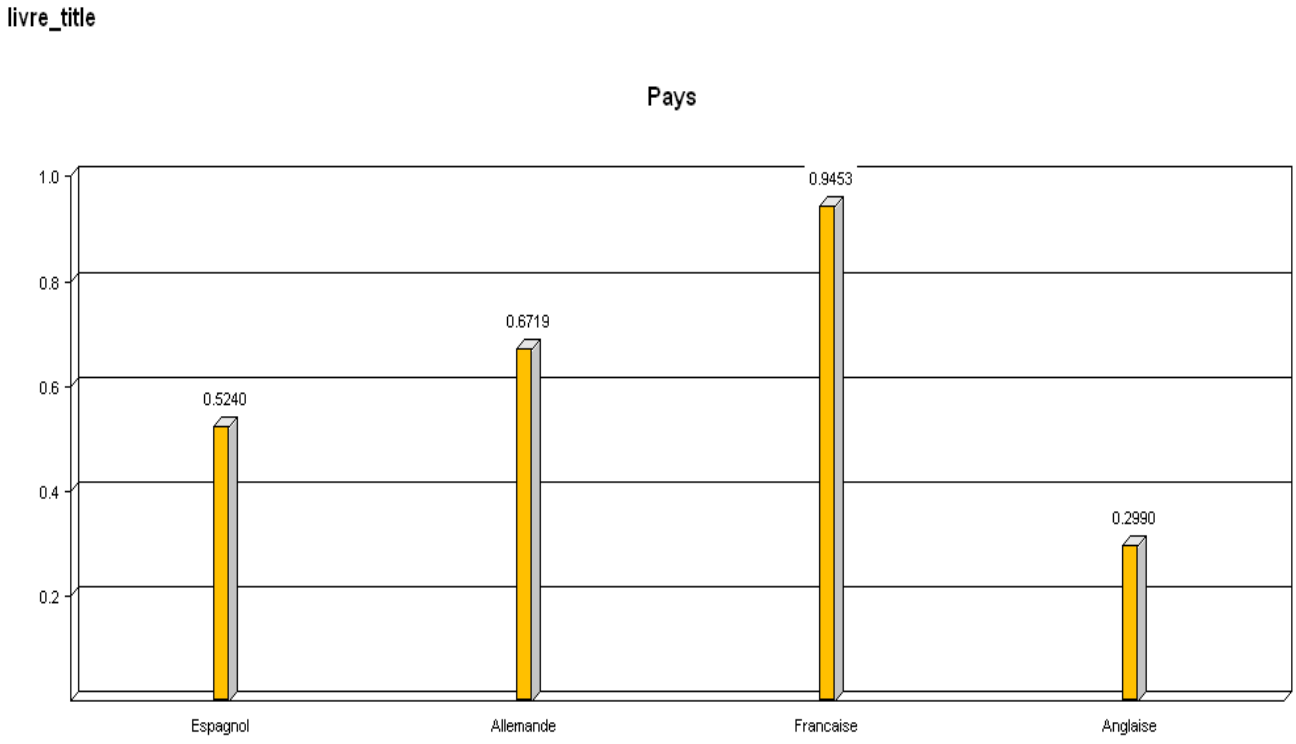
La méthode STAT, Elementary Statistics On Symbolic Objetcs, étend aux objets symboliques, représentés par leur description, plusieurs méthodes de statistique élémentaire limitées aux données.

Le choix de la méthode dépend du type de variable. Elle est filtrée en fonction de la méthode de travail :

- Fréquence relative pour les variables multimodales.
- Fréquence relative pour les variables intervalles.
- Capacité et min/max/moyenne pour variables multimodales probabilistes
- Biplot pour les variables multimodales.
- Objet central

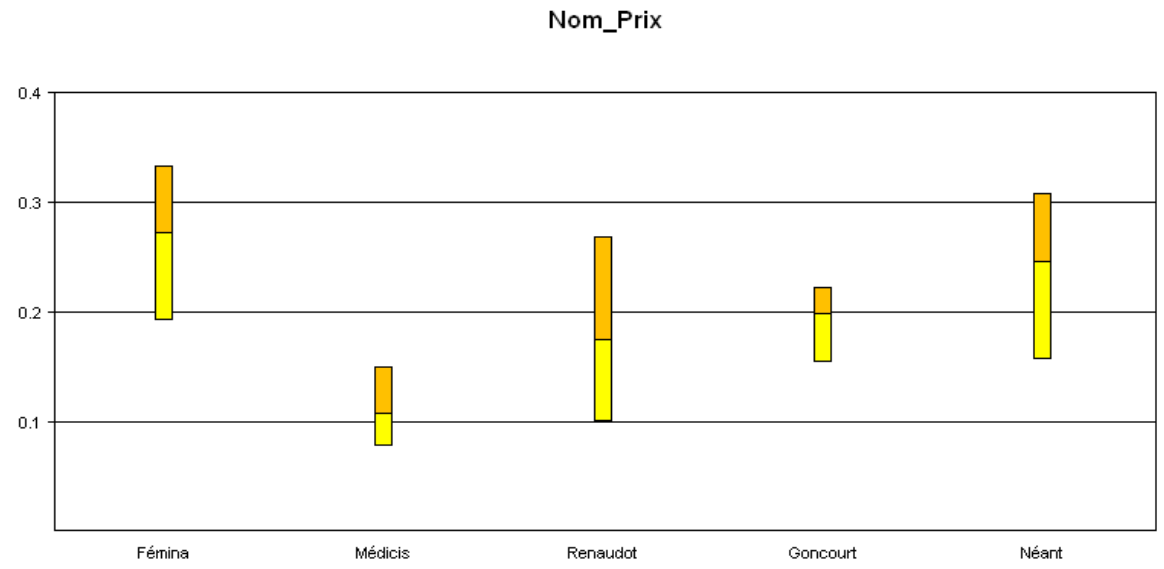
- Application

Nous avons choisis comme variable le pays de l'auteur.



Comme nous l'avons déjà remarqué précédemment, les auteurs des livres sélectionnés aux prix de littérature sont pour la plupart français, qu'il y a très peu d'auteur d'origine allemande dont les livres sont sélectionnés.

La variable sélectionnée est le nom du prix.



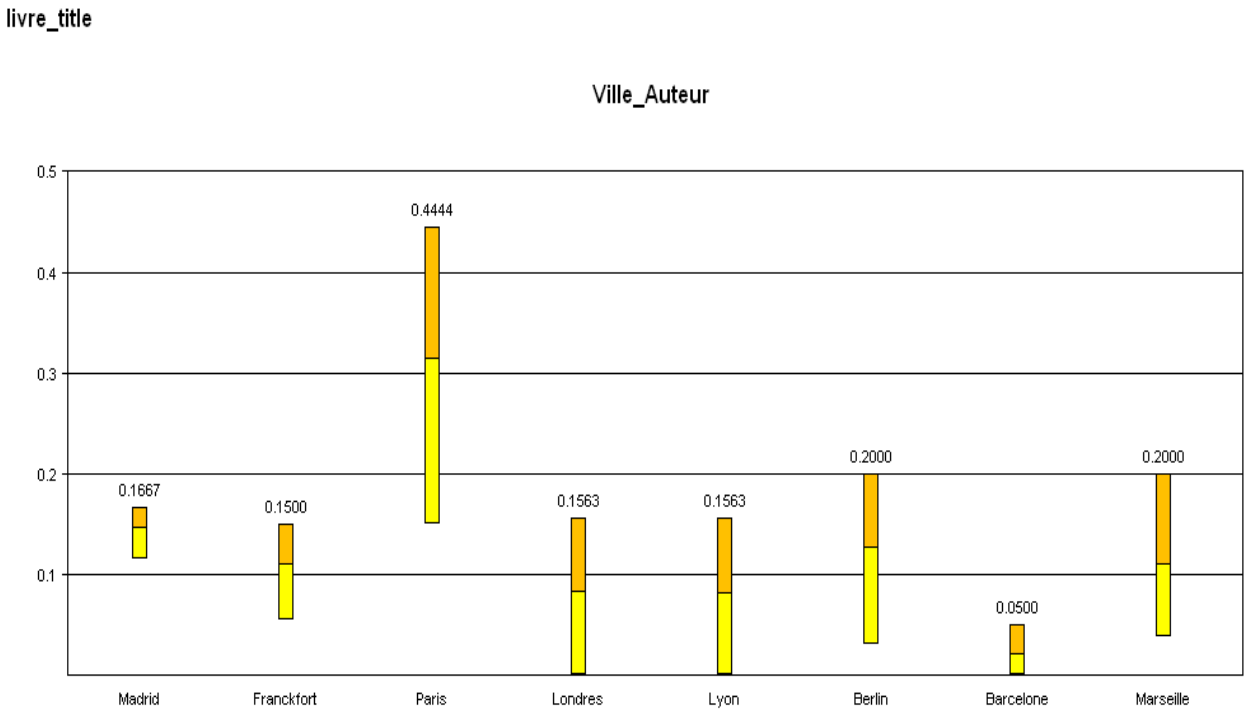
Min / mean / max - file LIVRE.SDS - variable 'Nom_Prix'

Nous remarquons que les moyennes les plus élevées sont le prix Fémina et aucun prix, on peut donc déduire que beaucoup de livres parmi les livres de la base n'ont reçu aucun prix car la moyenne est élevée.

Par contre, la moyenne des livres ayant reçu le prix Médicis est el plus faible, peu de livres ont reçu ce prix.

On peut donc conclure que la plus grande majorité des livres ont reçu un prix, bien que la plupart de ces livres primés ont reçu le prix Fémina le prix Renaudot a été aussi largement reçu par les livres sélectionnés.

La variable sélectionnée est la ville d'origine de l'auteur.



Selon le graphe, les livres sélectionnés au prix de littérature sont pour la grande majorité, des livres dont les auteurs sont originaires de la ville de Paris, étant donné la moyenne très élevé de Paris. Les auteurs originaires de la ville de Barcelone ont la moyenne la plus faible donc très peu d'auteur originaire de Barcelone.

3. Méthode DIV

- Présentation

DIV pour Divisise clustering, c'est une méthode de classification hiérarchique qui débute avec tous les objets d'une classe et procède par divisions successives de chaque classe. A chaque étape une classe est divisée en deux classes selon une question binaire, qui induit le meilleur partage en deux classes.

- Application

Les variables utilisées sont

- ✓ l'âge des auteurs
- ✓ le nombre de prix pour chaque catégorie d'ouvrages.

Le nombre de classe choisit est de 3.

Voici le résultat obtenu :

```
VARIANCE OF THE SELECTED VARIABLES :
-----------------------------------
Age                                 : 11.187500
Nb_prix                             : 43.687500
-----------------------------------------------------


PARTITION IN 2 CLUSTERS :
-------------------------:


Cluster 1 (n=3) :

"Littérature espagnole" "Littérature anglaise" "Littérature allemande"

Cluster 2 (n=1) :

"Littérature française"

Explicated inertia : 66.894457



PARTITION IN 3 CLUSTERS :
-------------------------:


Cluster 1 (n=1) :

"Littérature espagnole"

Cluster 2 (n=1) :

"Littérature française"

Cluster 3 (n=2) :

"Littérature anglaise" "Littérature allemande"

Explicated inertia : 85.193622
```

```
THE CLUSTERING TREE :
---------------------

    - the number noted at each node indicates
      the order of the divisions
    - Ng <-> yes and Nd <-> no


              +----  Classe 1 (Ng=1)
              !
        !----2- [Age <= 57.750000]
        !     !
        !     +----  Classe 3 (Nd=2)
        !
  !----1- [Nb_prix <= 30.500000]
        !
        +----  Classe 2 (Nd=1)
```

Premièrement, on note que 3 catégories de littérature ont obtenu moins de 30.5 prix en moyenne contre une catégorie qui a remporté plus de 30,5 prix.

Parmi les 3 catégories, on remarque que deux des catégories ont pour auteur, des personnes âgé de moins de 57 ans.

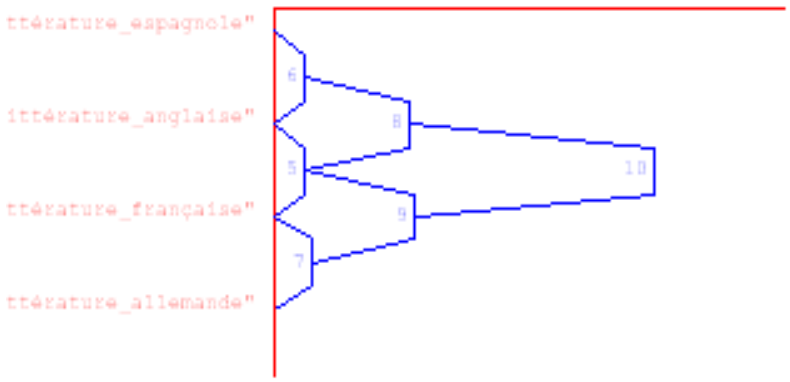
4. Méthode PYR

- Présentation

C’est une classification sous forme de pyramide. La pyramide est construite par un algorithme d’agglomération opérant par en bas, c'est-à-dire les concepts, vers le haut, en agglomérant les classes à chaque niveau supérieures.
Chaque classe est définie par son extension et également par un objet symbolique qui décrit les propriétés de la classe : l’intension de la classe.
L’intension est hérité d’un prédécesseur vers son successeur.

- Application

Les variables choisies sont : Nom d’auteur, prénom, pays, ville, âge, tranche d’âge, nom du prix, nombre de prix, éditeur.



Conclusion

SODAS est un logiciel qui offre une multitude de méthodes dans l'analyse de données. Grâce au logiciel, on ne travaille plus sur les individus de premier ordre mais sur des concepts des individus de premier ordre.

SODAS est un logiciel qui permet de synthétiser un nombre impressionnant de données d'informations dans le but d'extraire des informations pertinentes aux décideurs et ou chercheurs.

C'est également un logiciel assez simple d'utilisation, qui le rend accessible au plus novice. Avec des résultats pertinents et des représentations graphiques que l'on peut exploiter aisément et qui résume beaucoup d'informations.