1. Supports et tutoriels

[REF 1] Rakotomalala R., « Classification ascendante hiérarchique - Diapos » ;
http://tutoriels-data-mining.blogspot.fr/2016/07/classification-ascendante-hierarchique.html

[TUTO 1] Rakotomalala R., « Classification automatique sous R » ;
http://tutoriels-data-mining.blogspot.fr/2015/10/classification-automatique-sous-r.html

[TUTO 2] Rakotomalala R., « Interpréter la valeur test » ;
http://tutoriels-data-mining.blogspot.fr/2008/04/interprter-la-valeur-test.html

[TUTO 3] Rakotomalala R., « La complémentarité CAH et ACP » ;
http://tutoriels-data-mining.blogspot.fr/2008/03/la-complmentarit-cah-et-acp.html

[TUTO 4] Rakotomalala R., « Classification automatique sous Python » ;
http://tutoriels-data-mining.blogspot.com/2016/03/classification-automatique-sous-python.html

2. Outils – Excel + Tanagra

1. Chargez et installer le logiciel Tanagra sur votre ordinateur
http://chirouble.univ-lyon2.fr/~ricco/tanagra/fr/tanagra.html

2. Intégrez Tanagra dans Excel en tant que macro complémentaire
http://tutoriels-data-mining.blogspot.fr/2010/08/ladd-in-tanagra-pour-excel-2007-et-2010.html

3. Données

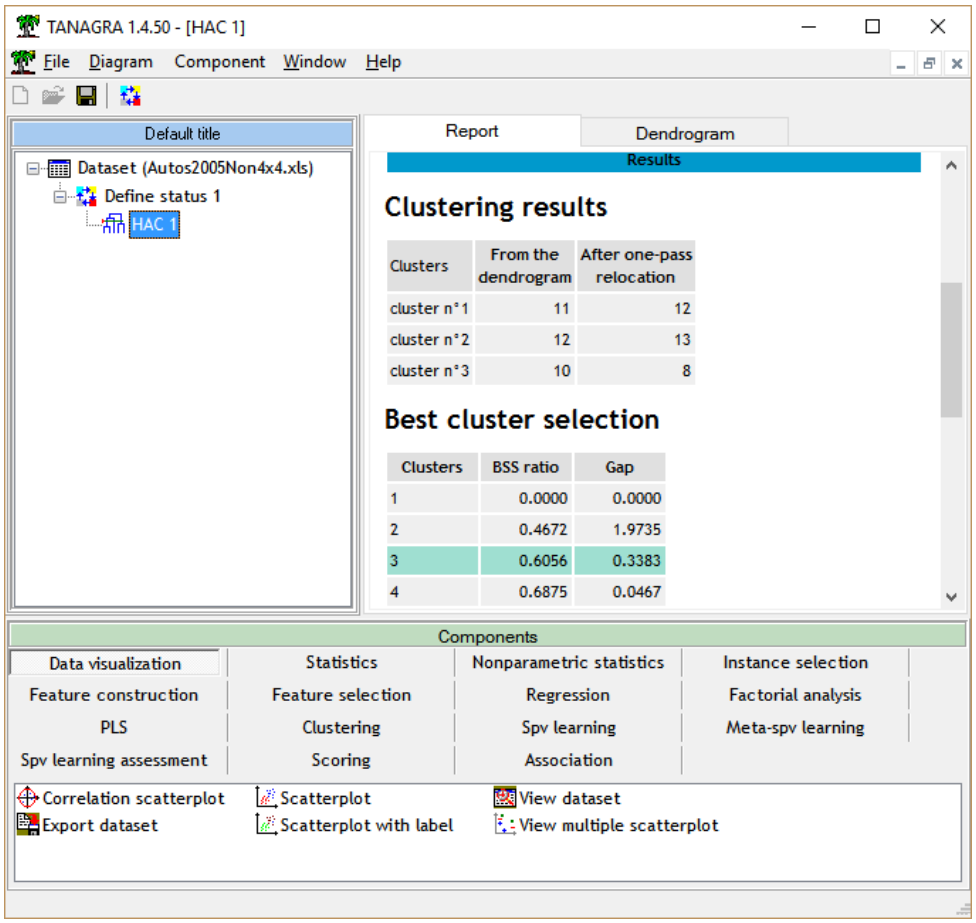
Nous utilisons les données « **Autos2005Non4x4.xls** » décrivant les caractéristiques de n = 33 véhicules. Il s'agit du fichier AUTOS2005.TXT, utilisé dans un des exercices pour l'analyse factorielle, pour lequel nous avons retranchés les véhicules étiquetés « 4 x 4 ».

- MODELE est un label non utilisé pour les calculs, mais sera utile pour l'interprétation des résultats ;
- **PUISSANCE … POIDS** sont les variables actives ;
- CO2 et PRIX sont les variables illustratives quantitatives ;
- ORIGINE et CARBURANT sont les illustratives qualitatives.

4. Modélisation et interprétation

A rendre : un fichier Excel avec le détail des calculs dans différentes feuilles du classeur que vous numéroterez suivant les questions. Lorsque le résultat est relatif à Tanagra, faites une copie d'écran que vous collez dans la feuille dédiée, insérez votre commentaire dans une zone de texte.

1. En suivant [TUTO 3, page 1 à 3], importez « **Autos2005Non4x4.xls** » dans Tanagra, définissez les variables INPUT de l'étude (**PUISSANCE … POIDS)**, puis lancez la modélisation avec le composant HAC. Vous devriez obtenir le résultat suivant :



2. La partition en 3 groupes suggérée par Tanagra vous paraît-elle justifiée si l'on considère le dendrogramme ?
3. Insérez le composant VIEW DATASET (onglet DATA VISUALIZATION) à la suite du HAC 1. Notez la présence de la colonne CLUSTER_HAC_1 générée automatiquement par Tanagra. Copiez (menu COMPONENT / COPY RESULTS) le tableau de données incluant cette nouvelle colonne dans une nouvelle feuille du classeur Excel.
4. **Nous travaillons sous Excel à partir de maintenant**. Calculez la distance au barycentre global des centres de classes pour les situer. Utilisez le carré de la distance euclidienne pondérée par l'inverse de la variance [REF 1, page 14]. Remarque : un tableau de calcul comme ci-dessous devrait vous permettre de parvenir à vos fins (*cf. utilisation des tableaux croisés dynamiques*)

	Données						
Cluster_HAC_1	Moyenne de puissance	Moyenne de cylindree	Moyenne de vitesse	Moyenne de longueur	Moyenne de largeur	Moyenne de hauteur	Moyenne de poids
c_hac_1	81.75	1442.83	168.67	378.75	167.75	151.83	1070.67
c_hac_2	165.38	2014.08	215.77	448.23	178.23	144.15	1399.00
c_hac_3	229.63	3243.75	232.88	486.75	186.88	151.13	1747.13
Total général	150.55	2104.45	202.79	432.30	176.52	148.64	1364.00

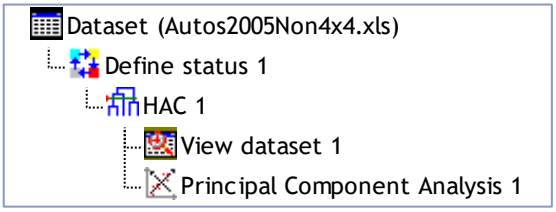
Ecart-typep	70.51	912.41	31.29	46.47	9.26	8.06	294.43

Vous devriez obtenir $d^2(C1) = 6.04$; $d^2(C2) = 0.70$; $d^2(C3) = 8.15$. Que peut-on en conclure ?

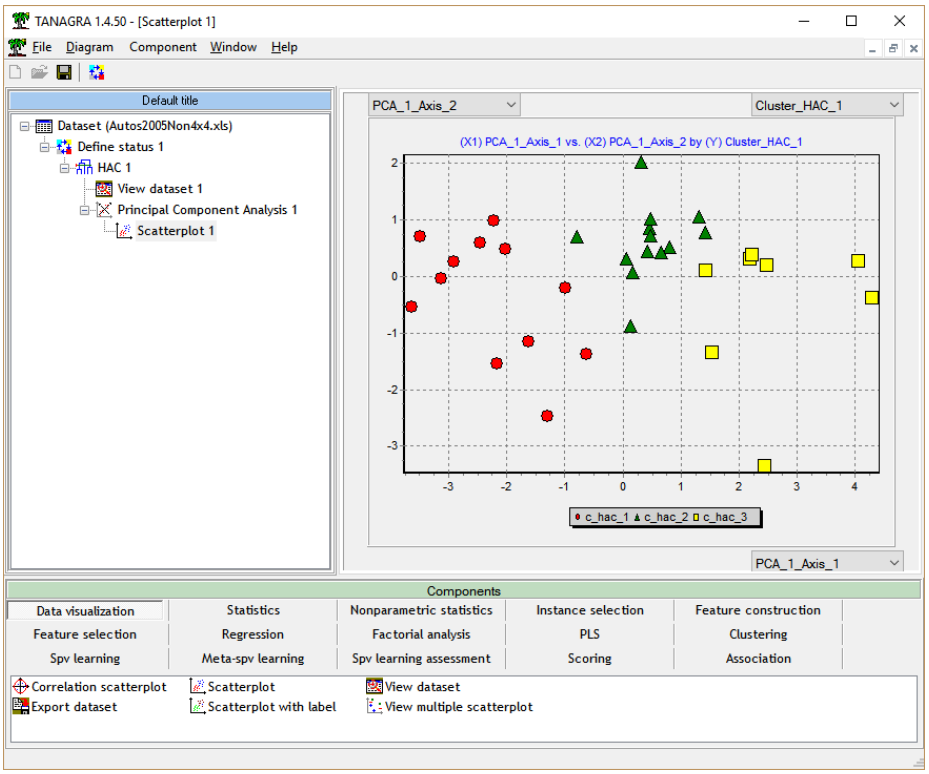
5. Calculez la distance entre les centres de classes pris deux à deux cette fois-ci, que concluez-vous ?

6. **Nous souhaitons caractériser les groupes à l'aide des variables**. Calculez le rapport de corrélation des variables quantitatives actives et illustratives au regard des groupes issus de la typologie. Quelles sont les variables qui permettent de distinguer au mieux les classes ?
7. Pour approfondir cette analyse, calculez les valeurs tests de chaque variable pour chaque classe [TUTO 2, pages 1 et 2]. Comment lire les résultats ?
8. Croisez les classes avec les variables qualitatives illustratives et calculez le V de Cramer. Quelles interprétations peut-on en tirer ?
9. **Intéressons-nous aux individus maintenant**. Pour chaque individu, calculez (d^2.own) le carré de la distance au barycentre de sa classe d'appartenance (avec le carré de la distance euclidienne pondérée par la variance toujours). Quels sont les 3 individus les plus représentatifs – les parangons – de chaque classe ? Cela conforte-t-il l'interprétation issue de l'analyse des variables réalisée précédemment ?
10. Pour chaque individu, calculer (d^2.next.closest) le minimum de la distance aux autres centres de classes (autres que sa classe affectée). Formez alors la quantité : $R^2 = 1 - d^2.own/d^2.next.closest$. Les individus bien représentés dans la classe ont un R^2 proche de 1 ; ceux qui posent problème ont un R^2 proche de 0, voire négatif. Dans chaque classe, quel est l'individu qui est le plus mal représenté ? Quels commentaires ces résultats vous inspirent-ils ?
11. **Nous revenons dans Tanagra**. Pour confirmer ou infirmer vos idées, réalisez une ACP à la suite de HAC [Voir http://tutoriels-data-mining.blogspot.fr/2008/03/acp-description-de-vhicules.html]. Est-ce que 2 facteurs suffisent pour représenter correctement les données ?

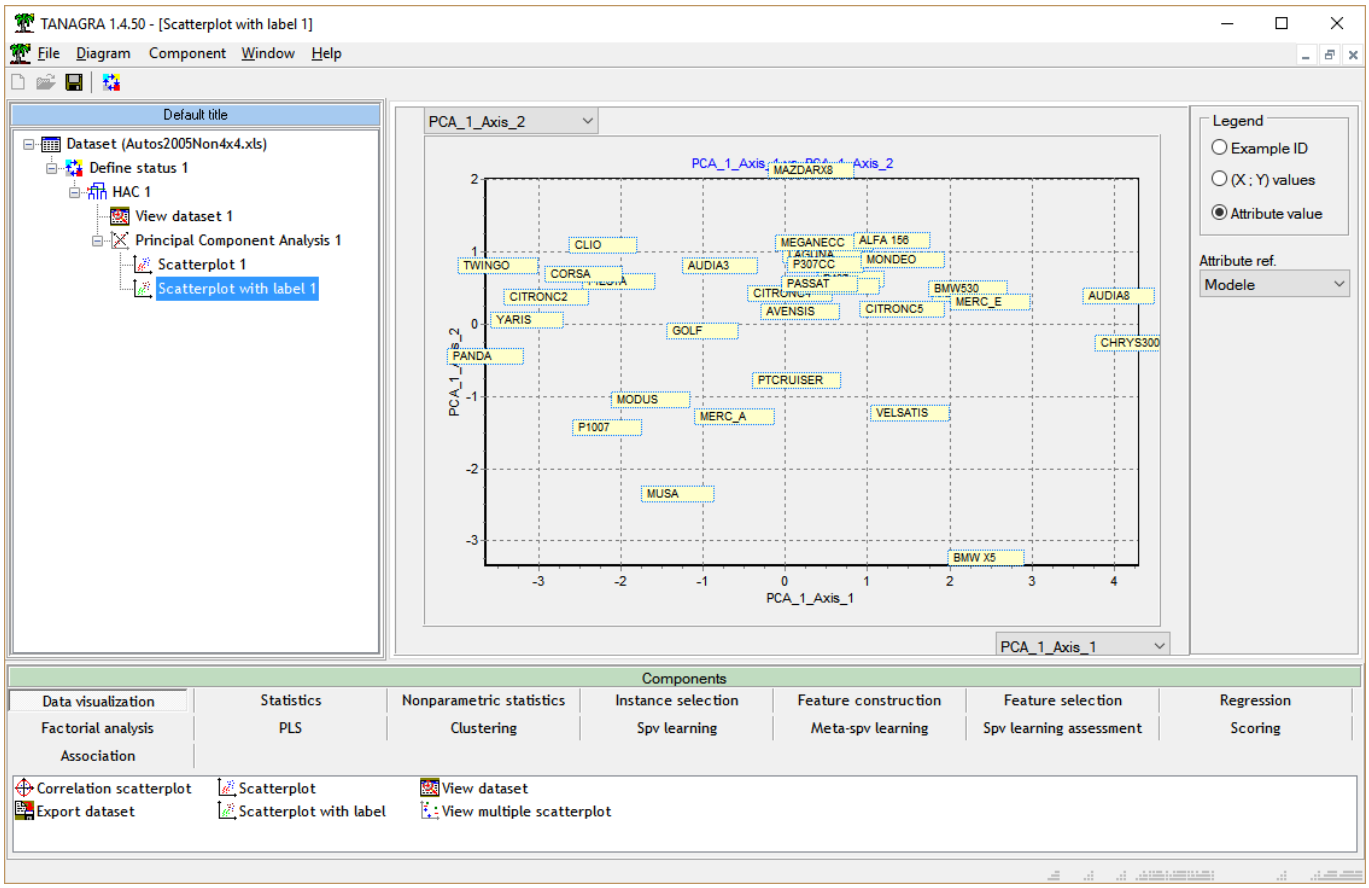
    A ce stade, votre diagramme devrait se présenter comme suit :



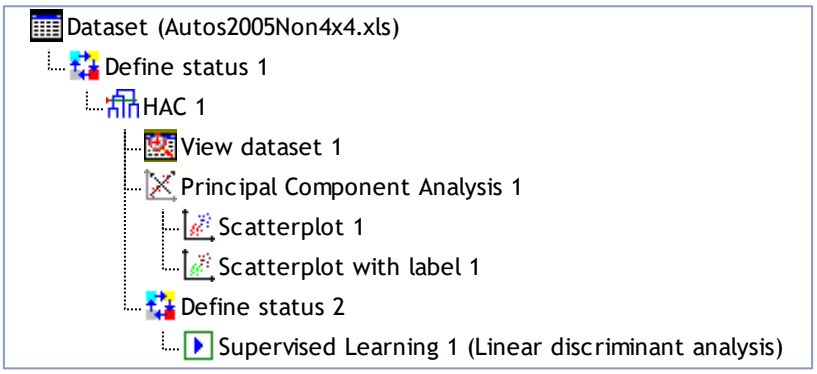
12. (*La réponse est OUI. Deux axes conviennent*). A l'aide du tableau des corrélations, dégagez les déterminants de chaque facteur. Quelles sont les propriétés des véhicules mis en exergue sur le premier facteur ? Sur le second ?
13. Avec le composant SCATTERPLOT (onglet DATA VISUALIZATION), représentez les observations dans le premier plan factoriel en les coloriant selon leurs classes d'appartenance. Votre diagramme devrait ressembler à ci-dessous. Que vous inspire ce graphique ? Est-ce que ce résultat va dans le même sens que l'analyse univariée des variables réalisée précédemment ?



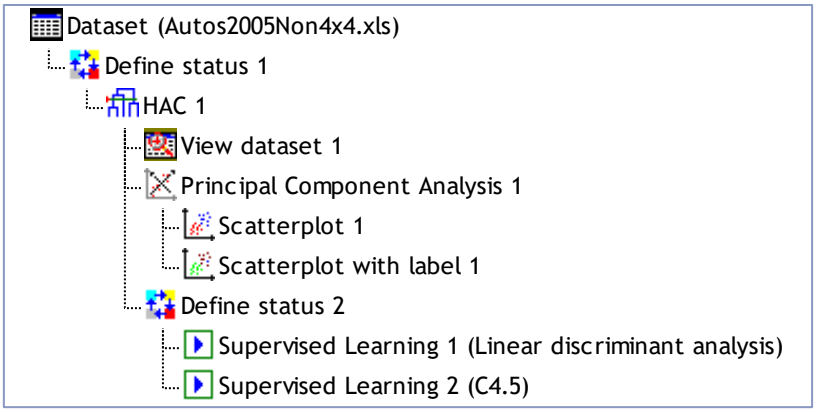
14. A l'aide du composant SCATTERPLOT WITH LABELS (onglet DATA VISUALIZATION), représentez les points dans le premier plan factoriel en leur associant leur label. Ce que nous voyons conforte-t-il les conclusions relatives à l'analyse des individus ?



15. Pour l'interprétation des groupes, mais aussi pour disposer d'une fonction d'affectation facile à manipuler, nous désirons mener une analyse discriminante linéaire (LINEAR DISCRIMANT ANALYSIS, onglet SPV LEARNING ; voir le thème « Data Mining 1 – Analyse prédictive) où la colonne des classes issues de la typologie fait office de variable cible, les variables actives de la typologie deviennent les variables explicatives. Insérez un composant DEFINE STATUS à la suite de HAC 1, mettez en TARGET la colonne CLUSTER_HAC_1, en INPUT les variables actives. Que constatez-vous à la lecture des résultats ? Les groupes sont-ils faciles à reconnaître ? Quelles sont les variables déterminantes finalement pour la désignation des groupes ? Votre diagramme devrait se présenter comme suit.



16. Pour conforter ce résultat assez étonnant (*si, si, ça l'est…*), nous souhaitons utiliser un arbre de décision dans l'explication des groupes. Nous utilisons la méthode C4.5 (onglet SPV LEARNING). Comment lire les règles issues de l'arbre ? Voici le diagramme au final :



Finalement, que faut-il penser de cette classification des véhicules ?

5. CAH sous R

Réitérez l'analyse complète sous R – en vous concentrant sur les calculs, il n'est pas nécessaire de refaire les interprétations – en utilisant la procédure **hclust()** du package STATS. Inspirez-vous de [TUTO 1] pour répondre aux questions.

A rendre : un rapport .docx issu d'un projet R Markdown.

Remarque : L'algorithme C4.5 est disponible (à peu de choses près) sous l'appellation J48 dans le package RWeka.

6. CAH sous Python

Réitérez encore une fois l'analyse sous Python en vous inspirant du [TUTO 4] cette fois-ci. Les algorithmes supervisés (analyse discriminante, arbre de décision) sont fournis par la librairie « scikit-learn ». L'affichage de l'arbre n'est pas trivial, cet exemple devrait vous aider : http://scikit-learn.org/stable/auto_examples/tree/plot_unveil_tree_structure.html

A rendre : un rapport PDF issu d'un notebook JUPYTER