# 1 Objectif

## Induction par arbre avec WinIDAMS 1.3.

WinIDAMS (**I**nternationally Developed **D**ata **A**nalysis and **M**anagement **S**oftware Package) est un logiciel de statistique développé sous l'égide de l'UNESCO. J'en avais dessiné les contours récemment (« Statistiques avec WinIDAMS », octobre 2014)[1]. J'avais remarqué durant cette étude la procédure SEARCH consacrée à l'apprentissage par arbre. Elle intègre les arbres de décision et de régression, ainsi qu'une méthode que l'on voit peu dans les logiciels, et qui s'apparente à une régression par morceaux. Plutôt que d'incorporer la description de SEARCH dans le document générique consacré à WinIDAMS, j'ai préféré écrire un tutoriel à part car plusieurs éléments avaient attiré mon attention.

(1) L'outil propose des sorties qui permettent de retracer le processus de construction de l'arbre. (2) Cette caractéristique est d'autant plus intéressante que la documentation technique décrit les formules utilisées avec force détail (« WinIDAMS Reference Manual (release 1.3) », april 2008 ; Chapter 56, « Searching for structure »). Nous pourrons ainsi de reproduire les calculs intermédiaires pour comprendre pleinement la teneur des méthodes. (3) J'avoue avoir été d'autant plus curieux d'étudier la procédure que j'avais remarqué parmi les contributeurs des auteurs qui ont énormément œuvré dans la popularisation de l'induction par arbre, notamment J.N. Morgan et J. Sonquist qui comptent parmi les références les plus anciennes et les plus prolifiques dans le domaine[2]. Mieux appréhender leur vision ne peut qu'améliorer notre compréhension de ces méthodes. (4) Enfin, la troisième option proposée par SEARCH (Analysis = Regression) correspond à une méthode que je n'ai jamais rencontré dans d'autres outils. Forcément, cela m'a interpelé. De par ma trajectoire scientifique, je suis toujours très curieux de tout ce qui touche aux arbres.

Ce tutoriel décrit les tenants et aboutissants des 3 options (CHI, MEANS, REGRESSION) de la procédure SEARCH de WinIDAMS[3,4]. Les numéros de page indiqués dans ce document font référence au manuel en anglais de WinIDAMS au format PDF accessible sur le web[5].

---

[1] http://tutoriels-data-mining.blogspot.fr/2014/10/statistiques-avec-winidams.html

[2] A une époque où les publications scientifiques ne se résumaient pas à une course aux « impact factor »…

[3] Une description de la procédure en français est accessible sur le web (options, exemples d'utilisation, lecture des résultats, restrictions, …) : http://www.unesco.org/webworld/portal/idams/html/french/F1search.htm

[4] Un document avec des exemples d'utilisation et la lecture des sorties de SEARCH est également accessible : http://www.unesco.org/webworld/idams/advguide/ex10.htm

[5] http://portal.unesco.org/ci/en/files/7671/12155211213ManualR13E.pdf/ManualR13E.pdf

# 2 Arbre de décision

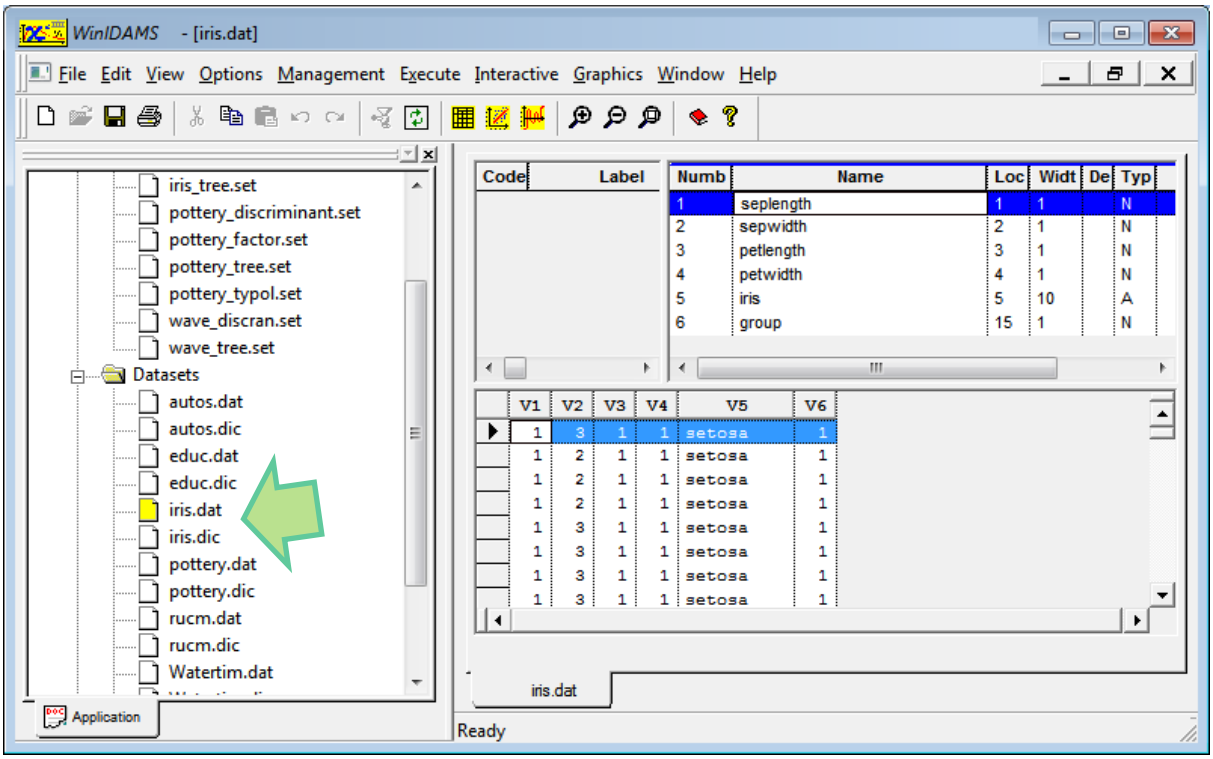
## 2.1 Objet de la méthode

L'option ANALYSIS = CHI de SEARCH traite le cas d'une variable cible qualitative (section 56.3). Les variables prédictives peuvent être quelconques. A l'usage je me suis rendu compte que la procédure ne traite que des variables discrètes codées (1, 2, 3…)[6]. Elles peuvent être nominales ou ordinales, et traitées en conséquence durant le processus de binarisation c.-à-d. selon le cas, l'algorithme peut regrouper des modalités non-contigües ou non.

SEARCH sait appréhender les problèmes à plusieurs variables cibles. Il faut qu'elles soient binaires, les combinaisons sont traitées comme des modalités d'une variable qualitative. Par exemple, une variable G à 4 modalités définie dans {(1,1), (1,0), (0,1), (0,0)} est dérivée de 2 variables cibles binaires $Y_1$={1,0} et $Y_2$={1,0}.

## 2.2 Données

Nous utilisons une version discrétisée[7] du fameux fichier IRIS[8]. Voici la description (iris.dic) et les premières observations (iris.dat) dans l'environnement WinIDAMS.[9]




[6] Mes tentatives d'utilisation de variables continues ont échoué en tous les cas. Peut être qu'une option m'a échappé.

[7] Les variables (V1 à V4) ont été recodées en qualitatives ordinales via un processus de découpage en intervalles.

[8] http://en.wikipedia.org/wiki/Iris_flower_data_set

[9] Voir http://tutoriels-data-mining.blogspot.fr/2014/10/statistiques-avec-winidams.html pour l'importation des données.

L'objectif est de prédire, à partir des descripteurs (V1 à V4), les valeurs prises par la variable cible « iris » définie dans {setosa, versicolor, virginica}. Par commodité, « iris » a été recodée en « group » prenant ses valeurs dans {1, 2, 3}.

## 2.3 Définition des traitements

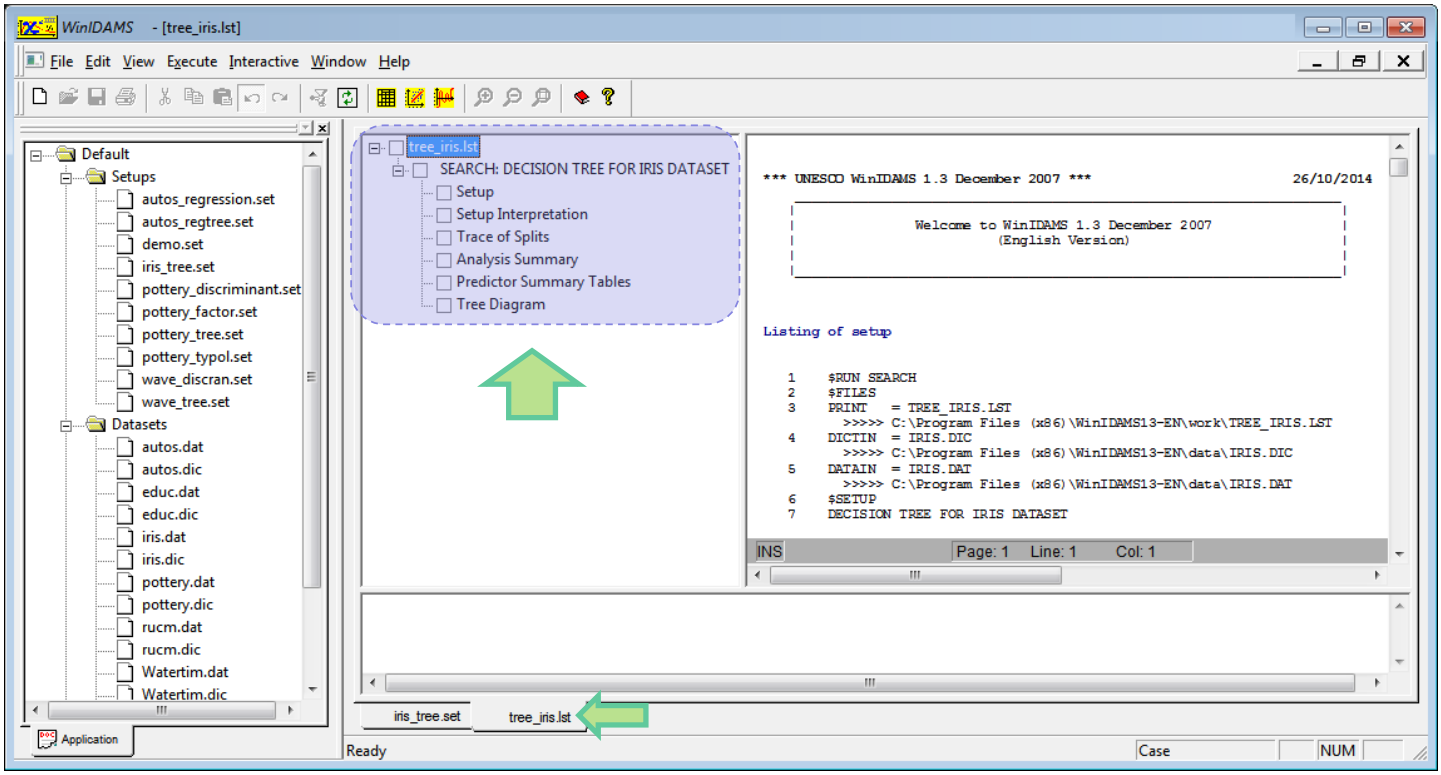
Nous devons créer un fichier « setup » pour définir les traitements à réaliser. Nous faisons appel à la procédure SEARCH.

```
$RUN SEARCH
$FILES
PRINT   = tree_iris.lst
DICTIN  = iris.dic
DATAIN  = iris.dat
$SETUP
DECISION TREE FOR IRIS DATASET
ANALYSIS=CHI  -
DEPV=V6 CODE=(1-3) -
MINC=10 -
PRINT=(FINAL,TREE,TRACE)
VARS=(V1-V4) TYPE=M
```

Voyons ce qu'il en est des options de l'étude :

- ANALYSIS spécifie le type d'analyse. Nous souhaitons construire un arbre de décision, avec une variable cible catégorielle (ANALYSIS = CHI).
- La variable cible est GROUP (DEPV = 6ème variable), à 3 modalités CODE=(1-3).
- Un groupe – une feuille issue d'une opération de segmentation – n'est validée que si elle couvre au moins MINC = 10 observations.
- Nous incluons dans les sorties : les tableaux récapitulatifs (FINAL), le dessin de l'arbre (TREE) et le détail des étapes de construction du modèle (TRACE).
- Les 4 premières variables de la base [VARS=(V1-V4)] constituent les variables prédictives. Elles seront traitées comme des variables ordinales (TYPE = M c.-à-d. monotonic). De fait, seules les modalités adjacentes seront regroupées lors de la recherche des meilleures segmentations binaires.

A l'issue des traitements, les sorties (fichier « **tree_iris.lst** ») sont découpées en plusieurs sections comme nous pouvons le voir dans la copie d'écran ci-dessous.




## 2.4 Lecture des résultats

### 2.4.1 Dessin de l'arbre

Les sommets sont numérotés (group 1, 2, 3, 4 et 5).

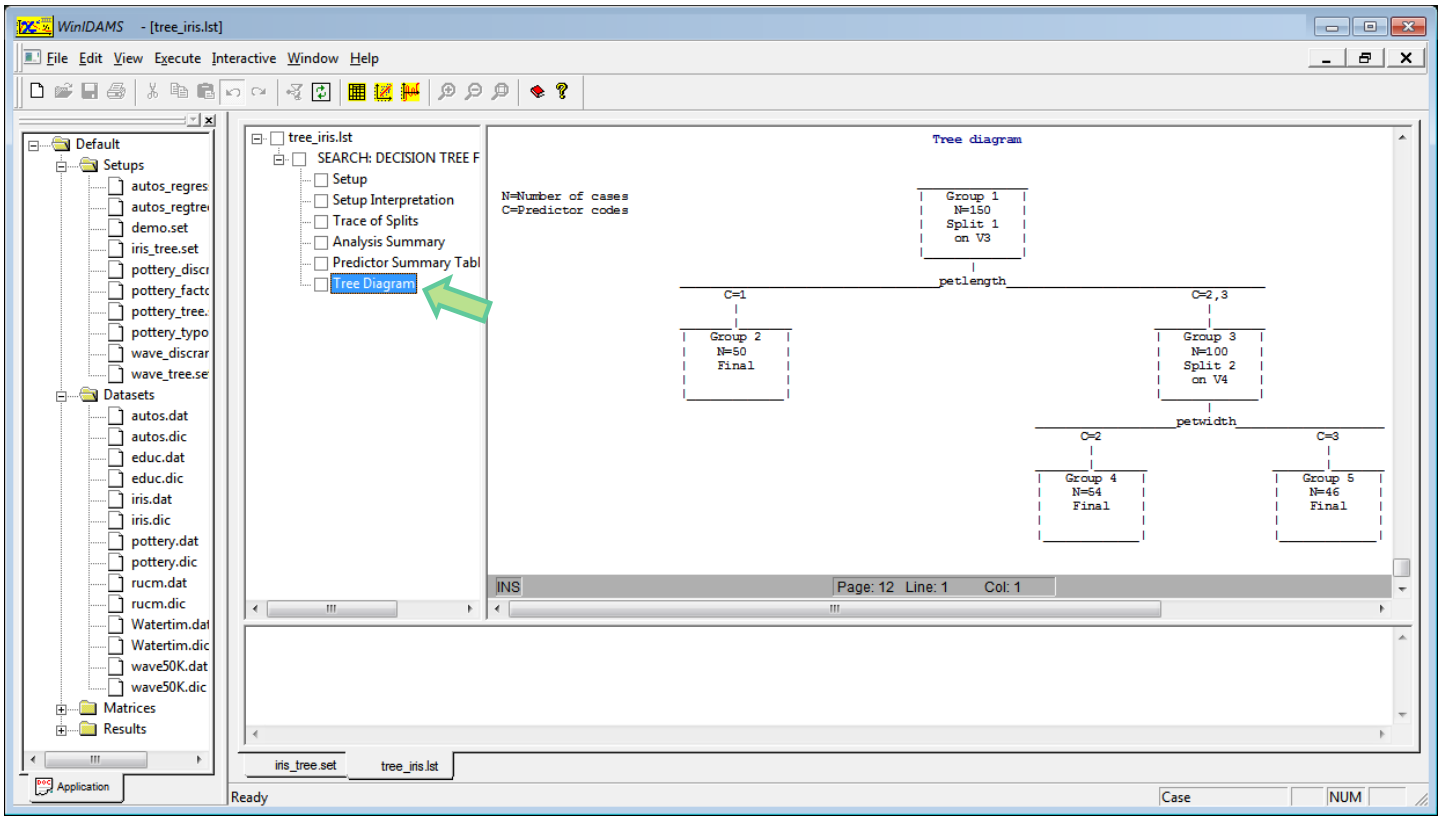



**Figure 1 - Représentation "graphique" de l'arbre de décision - WinIDAMS**

Nous obtenons un arbre à 3 feuilles avec respectivement 50 (group 2), 54 (group 4) et 46 (group 5) observations. La variable « petal-length » (V3) a été utilisée pour segmenter la racine (group 1), avec la modalité 1 d'un côté, les modalités 2 et 3 de l'autre. Puis, le group 3 a été partitionné avec « petal-width » (V4), opposant les modalités 2 et 3 (il n'y avait aucun individu de la modalité 1 dans le group 3).

### 2.4.2 Distributions dans les groupes

Dans la section « Final Group Summary Table » du rapport, nous observons les effectifs par sommet ; les distributions des classes sont dans « Percent distribution of the dependent variable for each group ».

```
Final group summary table

Group   2          50 cases                          Variation=0.00000000E+00
Group   4          54 cases                          Variation=0.33317509E+02
Group   5          46 cases                          Variation=0.96353846E+01


Percent distribution of the dependent variable for each group (*=Final g

                1       2*        3        4*       5*

Code= 1     33.33   100.00     0.00     0.00     0.00
Code= 2     33.33     0.00    50.00    90.74     2.17
Code= 3     33.33     0.00    50.00     9.26    97.83
```

Nous pouvons ainsi reconstituer les effectifs sur les différents sommets de l'arbre (Figure 2).

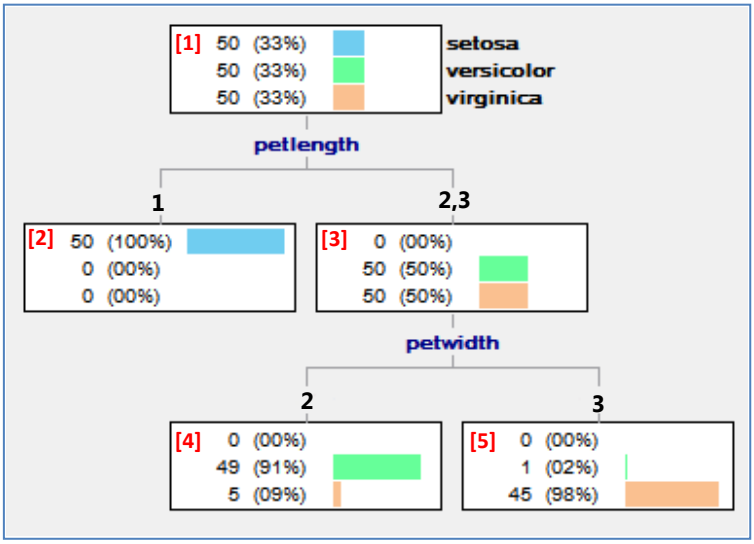



**Figure 2 - Représentation de l'arbre avec les effectifs et les distributions (SIPINA)**

Pour la modalité « versicolor » du sommet **n°4** par exemple, nous avons :

$$54 * 90.74 / 100 = 49$$

### 2.4.3 Dispersion (variation) dans les groupes

A chaque nœud est associé une mesure de dispersion des classes appelée « Variation ».

- Elle est nulle lorsque le groupe est pur, dans le sens où seule une classe – une modalité de la variable cible – est représentée. C'est le cas du groupe n°2 où les 50 individus qui s'y trouvent portent tous l'étiquette « setosa ».
- Elle prend sa valeur la plus élevée lorsque les classes sont équidistribuées c.-à-d. lorsque l'incertitude est totale.

Pour le groupe n°4 avec un effectif de 54 individus, la variation est obtenue avec (section 56.3, page 392) :

$$-2 * (0 * \ln 0 + 49 * \ln 0.9074 + 5 * \ln 0.926) = 33.3175$$

Le groupe 5 est plus pur dans le sens où la décision est plus tranchée avec une distribution de (0%, 2.17% et 97.83%). La variation est égale à 9.6354.

### 2.4.4 Partition finale – Tableau d'analyse de variance

La variation d'un groupe s'apparente à une variance conditionnelle. Leur addition équivaut à une variance intra-groupe c.-à-d. une variance résiduelle qui n'est pas expliquée par l'appartenance aux groupes. En suivant ce principe, WinIDAMS propose un tableau d'analyse de variance pour évaluer la qualité de la partition finale.

```
The partitioning ends with 3 final groups

The variation explained is 87.0%

One-way analysis of final groups

         Source          Variation          DF
         Explained 0.28663080E+03            2
         Error     0.42952911E+02          147
         Total     0.32958371E+03          149
```

La variabilité totale est calculée à partir de la racine :

$$SCT = -2 * (50 * \ln (50/150) + 50 * \ln (50/150) + 50 * \ln (50/150)) = 329.5837$$

La variabilité résiduelle est obtenue avec l'addition des variations conditionnelles

$$SCR = 0.000 + 33.3175 + 9.6354 = 42.9529$$

La variabilité expliquée par l'appartenance aux groupes est obtenue par différence

$$SCE = 329.5837 - 42.9529 = 286.6308$$

Nous pouvons ainsi déduire la part de variation expliquée par l'arbre de décision

$$R^2 = SCE / SCT = \mathbf{87\%}$$

### 2.4.5 Processus de construction

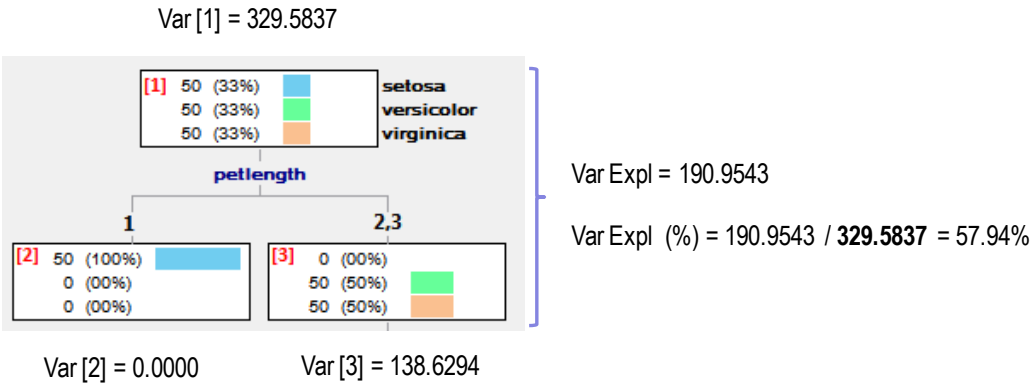
Avec l'option (PRINT = TRACE), WinIDAMS fournit les informations permettant de reconstituer les choix de segmentation sur les sommets. Prenons le premier sommet (n°1). Nous disposons des éléments suivants :

```
Split   1     candidate groups

    Group       N    Sum(WT)    Variation
       1     150 0.15000E+03 0.32958E+03

#Variation initiale du sommet n°1 à traiter
Attempt to split group   1       Var= 329.58371

#Performances des prédicteurs candidats
Predictor V1      seplength                Rank 1  Type M
  Codes  1 2 3
  Best split after code  1  Var expl=0.11587331E+03

Predictor V2      sepwidth                 Rank 1  Type M
  Codes  1 2 3
  Best split after code  2  Var expl=0.55710625E+02

Predictor V3      petlength                Rank 1  Type M
  Codes  1 2 3
  Best split after code  1  Var expl=0.19095427E+03

Predictor V4      petwidth                 Rank 1  Type M
  Codes  1 2 3
  Best split after code  1  Var expl=0.19095427E+03

#Choix de la meilleure variable de segmentation
Best split for group   1 on predictor V3      petlength                 Rank 1
     Var expl=0.19095427E+03

#Conformation de la partition (1) à gauche, (2,3) à droite.
Split group   1 on V3      petlength                  Var expl=0.19095427E+03
    Into group   2, codes  1
     and group   3, codes  2 3
```

« Petal Length » est sélectionnée avec une variation expliquée de **190.9543**. « Petal Width » est aussi performante. Le logiciel a choisi arbitrairement la première.

La variation expliquée (190.9543) est obtenue par la différence entre la variation du sommet initial (n°1) (329.6857) et la somme des variations calculées sur les sommets (n°2 et 3) issus de la segmentation (0 + 138.6294).

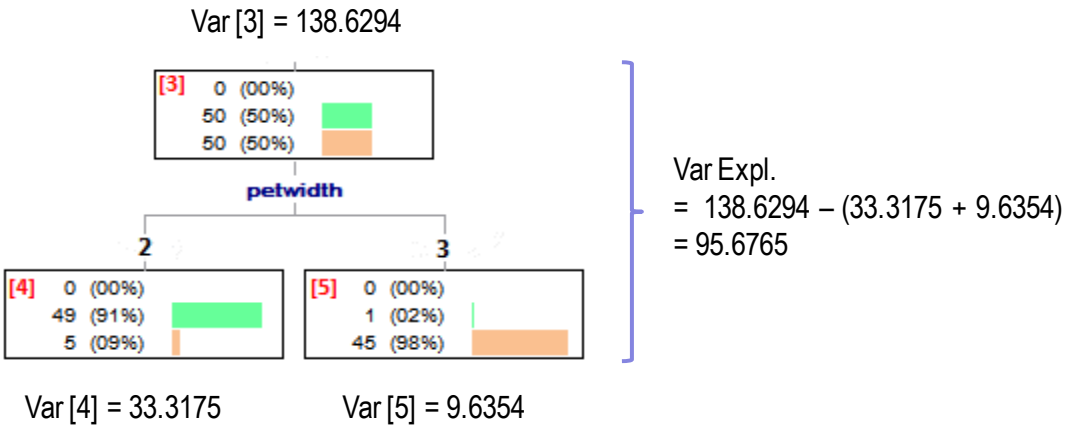
Les calculs sont résumés dans le graphique suivant.




Pour la segmentation du sommet n°3, WinIDAMS nous indique :

```
#Variation du sommet à traiter
Attempt to split group   3       Var= 138.62944

#Performances des variables candidates
 Predictor V1      seplength                 Rank 1  Type M
   Codes  1 2 3
   Best split after code  2  Var expl=0.22250025E+02

 Predictor V2      sepwidth                  Rank 1  Type M
   Codes  1 2 3
   Best split after code  1  Var expl=0.69116211E+01

 Predictor V3      petlength                 Rank 1  Type M
   Codes  2 3
   Best split after code  2  Var expl=0.91131355E+02

 Predictor V4      petwidth                  Rank 1  Type M
   Codes  2 3
   Best split after code  2  Var expl=0.95676544E+02

#Choix de la meilleure variable de segmentation
 Best split for group   3 on predictor V4      petwidth                  Rank 1
      Var expl=0.95676544E+02

#Configuration du découpage
 Split group   3 on V4      petwidth                   Var expl=0.95676544E+02
     Into group   4, codes  2
      and group   5, codes  3
```

« Petal width » est sans conteste le meilleur prédicteur ici, avec une variation expliquée de 95.6765. Voyons le détail des opérations :




**Rapportée à la variation initiale**, cela représente (95.6765 / **329.5837**) = 29.03%.

La construction est stoppée à ce stade, aucune nouvelle segmentation n'étant éligible. La part de variation expliquée par l'arbre dans son ensemble est égale à :

**57.94 % + 29.03 % = 86.97 %**

### 2.4.6 Impact global des variables

WinIDAMS indique l'impact des variables dans l'ensemble des segmentations effectuées ou envisagées dans l'arbre.

```
Per cent of total variation explained by best split for each group (*=Final groups)

                1       2*        3       4*       5*
 V1         35.16     0.00     6.75     0.03     0.00
 V2         16.90     0.00     2.10     0.08     0.28
 V3         57.94     0.00    27.65     0.00     0.00
 V4         57.94     0.00    29.03     0.00     0.00
```

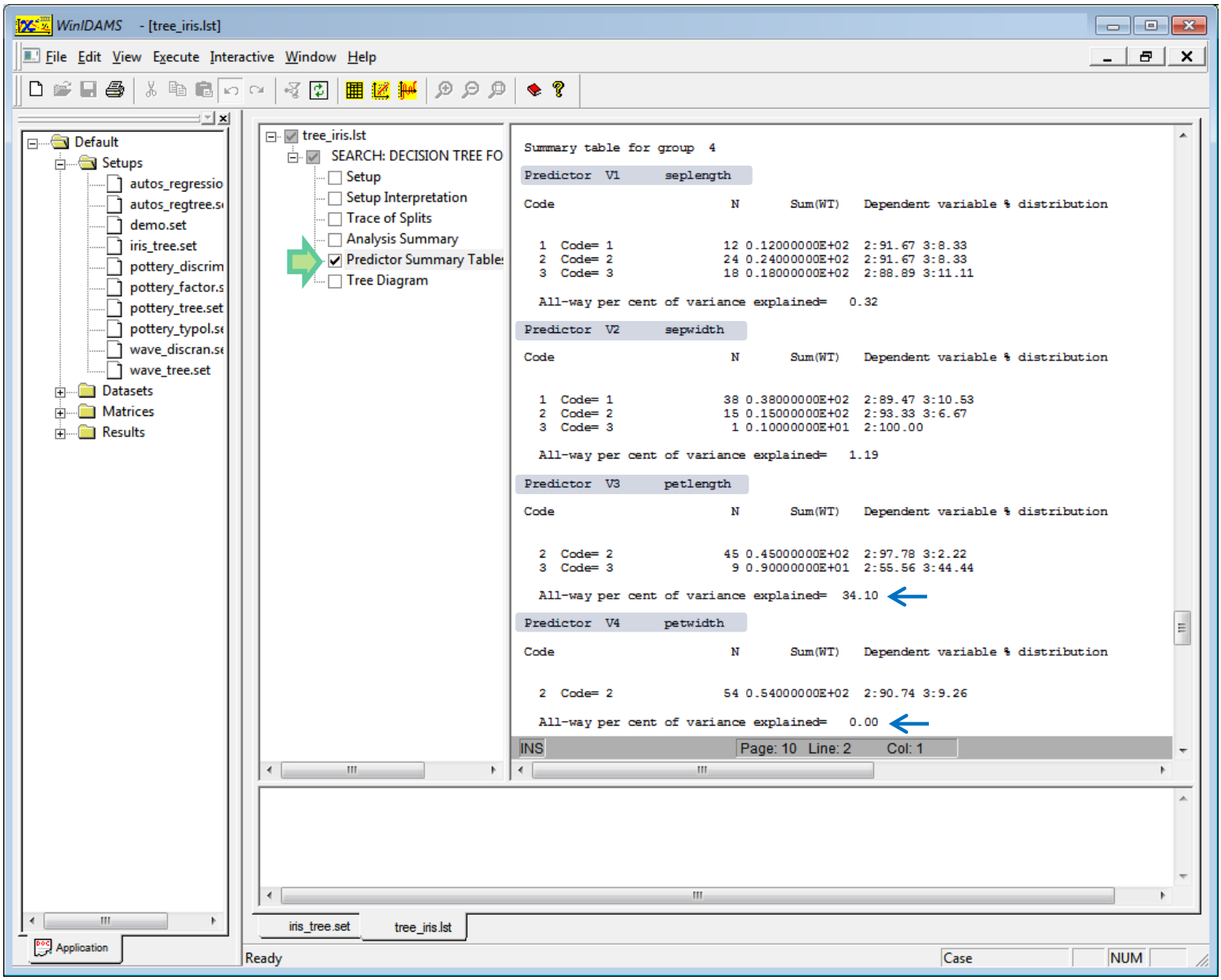
Ce tableau permet de mieux situer le rôle que peut jouer chaque prédicteur dans l'explication de la variable cible. En effet, s'en tenir au seul arbre induit (Figure 1) est trop réducteur. Certaines variables sont influentes mais peuvent être masquées par celles finalement sélectionnées lors des différentes segmentations. N'oublions que la première place peut parfois tenir à très peu de choses (ex. V3 et V4 étaient équivalentes lors du traitement du sommet n°1, V3 a été sélectionnée de manière arbitraire). Manifestement, les deux dernières variables (petal length et petal width) se démarquent nettement dans la prédiction de la catégorie d'iris. Elles présentent un impact globalement quasi-équivalent.

### 2.4.7 Segmentations candidates sur les feuilles

La construction de l'arbre a été stoppée parce qu'aucune segmentation candidate n'a été validée. Elles ne répondent pas aux conditions d'acceptabilité. Néanmoins, il peut être

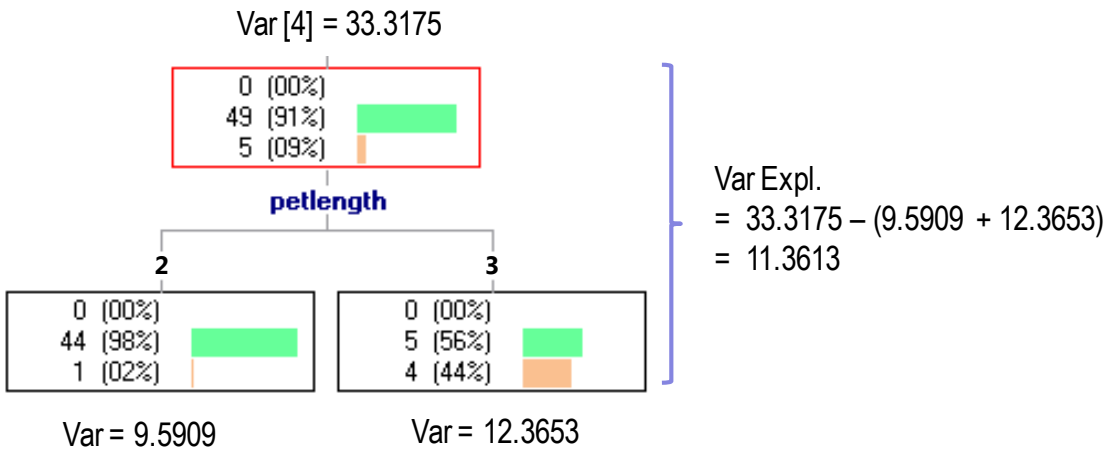
intéressant d'étudier les solutions proposées sur les feuilles, parce qu'elles peuvent s'avérer finalement pertinentes et inspirer de nouvelles analyses avec des paramétrages différents.

Voici ce que propose WinIDAMS pour l'éventuelle segmentation du sommet n°4 dans la section « Predictor Summary Table » du rapport :




L'outil fournit les segmentations candidates avant binarisation.

- « Petal width » est la variable la moins intéressante. Il n'y a pas de partitionnement possible, par conséquent la variation expliquée est nulle.
- « Petal length » est la plus performante. WinIDAMS fournit les effectifs des sous-groupes (45 individus dans le 1er sous-groupe, 9 dans le second), et les distributions des classes. Voici la représentation graphique de la segmentation candidate :




Très curieusement, le logiciel exprime la part de variance expliquée par rapport à la variation du sommet précédent (**34.10 % = 11.3613 / 33.3175**) et non pas par rapport à celle du sommet initial, comme il le fait dans les autres tableaux. Certes, cela n'a pas d'incidence sur le classement des variables sur un sommet. Mais la tentation était grande de confronter le chiffre avec ceux des segmentations déjà introduites (57.94% et 29.03%). Cette comparaison n'a pas lieu d'être puisque le référentiel n'est pas le même.

### 2.5 Conclusion

La procédure SEARCH est très complète. Nous disposons de tous les éléments pour expertiser finement le déroulement du processus de modélisation. Certaines sections du rapport, le tableau d'analyse de variance notamment, ouvrent la porte à un décryptage original des résultats. La construction de l'arbre s'inscrit dans un processus de décomposition de la variance. Le modèle peut s'évaluer en termes de proportion de variance expliquée. Si cette lecture est assez usuelle dans le cas d'une variable cible quantitative (comme nous le verrons dans la section suivante), elle l'est moins lorsqu'elle est qualitative. Nous constatons ici qu'elle est pourtant parfaitement licite[10].

## 3 Arbre de régression (I) – Arbre MEANS

### 3.1 Objet de la méthode

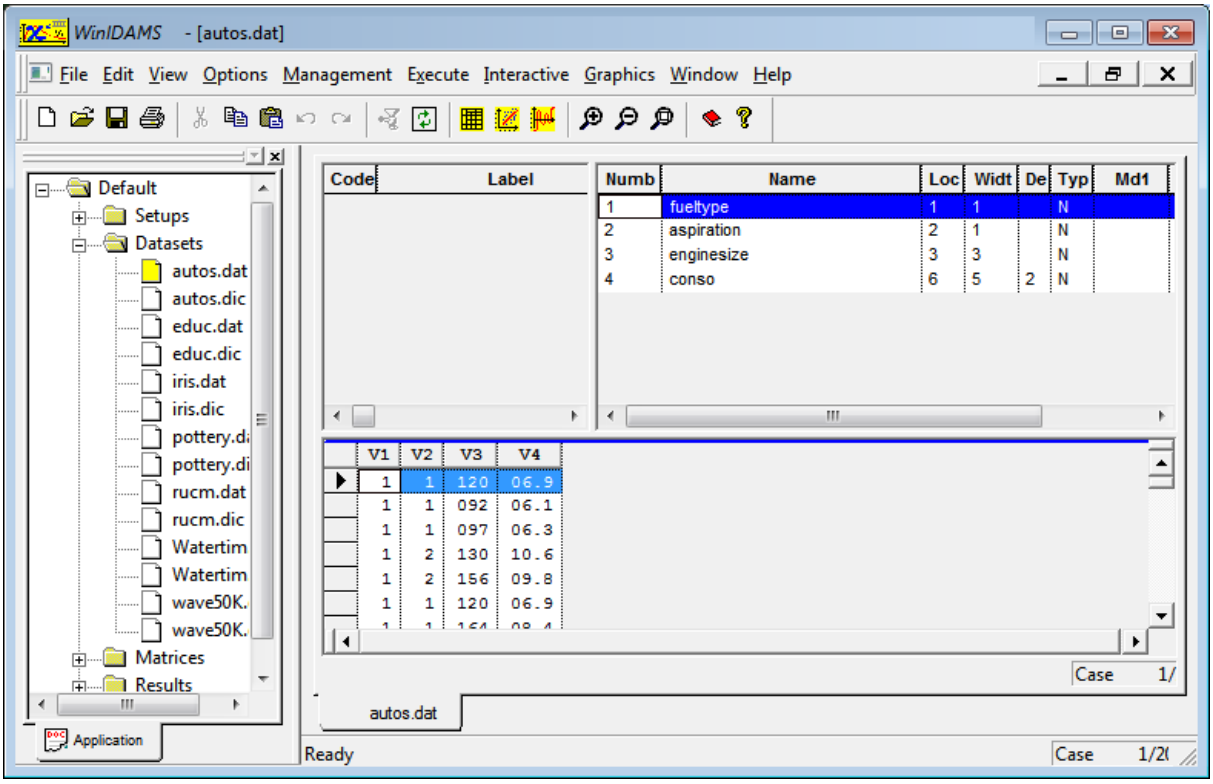
L'option ANALYSIS = MEANS de SEARCH traite des problèmes à variable cible quantitative et prédicteurs discrets. **On parle d'arbre de régression selon la terminologie usuelle en modélisation statistique** (Breiman et al., 1984 ; chapitre 8 « Regression Trees »). Pourtant, en

[10] Considérer l'arbre sous l'angle d'un processus de maximisation de la vraisemblance était une autre vision en vogue dans milieu des années 90. On pouvait ainsi s'appuyer sur des critères de type AIC (Akaike) ou BIC (Schwartz) pour sélectionner l'arbre « optimal ». Calculer la pénalité liée à la complexité de l'arbre n'était pas facile.

y regardant de plus près, l'appellation MEANS n'est pas dénuée de sens dans la mesure où la moyenne est la prédiction utilisée sur les nœuds de l'arbre, et que le processus de modélisation se réfère à une décomposition de la variance (qui est une mesure de dispersion autour de la moyenne). Pour éviter la confusion avec l'option présentée plus bas (section 4), nous parlerons d'arbre MEANS dans cette section.

## 3.2 Données

Nous utilisons une fraction du fichier AUTOS accessible sur le serveur UCI[11]. Ne sont conservées que les variables « fuel-type », « aspiration », « engine size » et « conso »[12]. Voici les premières lignes du fichier (autos.dat) décrit par le dictionnaire « autos.dic ».




Nous cherchons à prédire / expliquer la consommation (conso, quantitative) à partir du type de carburant (fuel type = {1 : gas, 2 : diesel}) et du mode d'alimentation (aspiration = {1 : std, 2 : turbo}) (discrètes).

## 3.3 Définition des traitements

Nous faisons appel à la procédure SEARCH avec l'option MEANS pour définit les traitements. Voici le fichier SETUP, plusieurs paramètres retiennent notre attention :

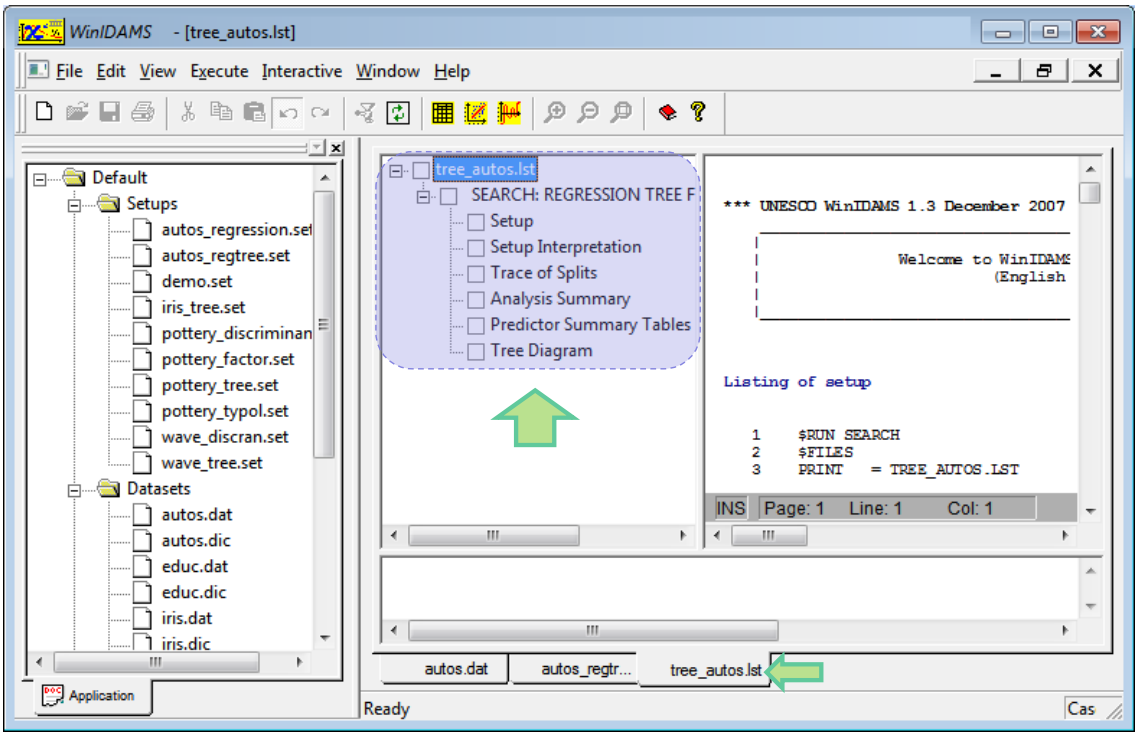
[11] https://archive.ics.uci.edu/ml/datasets/Automobile

[12] « Conso » est une conversion en L/100km de la variable « highway mpg » (en miles par gallon).

- ANALYSIS = MEANS spécifie le type d'analyse ;
- La variable cible est « conso » (DEPV = V4) ;
- Une segmentation est validée si toutes les feuilles comportent au moins 10 observations (MINC = 10) ;
- Les variables prédictives sont « fueltype » et « aspiration » [VARS = (V1-V2)], elles sont ordinales (TYPE=M). Cette dernière précision n'est pas déterminant dans notre exemple car V1 et V2 sont binaires, il n'y aura pas de regroupements lors des segmentations.

```
$RUN SEARCH
$FILES
PRINT   = tree_autos.lst
DICTIN  = autos.dic
DATAIN  = autos.dat
$SETUP
REGRESSION TREE FOR AUTOS DATASET
ANALYSIS=MEANS  -
DEPV=V4 -
MINC=10 -
PRINT=(FINAL,TREE,TRACE)
VARS=(V1-V2) TYPE=M
```

Ici également, les sorties sont décomposées en plusieurs sections.



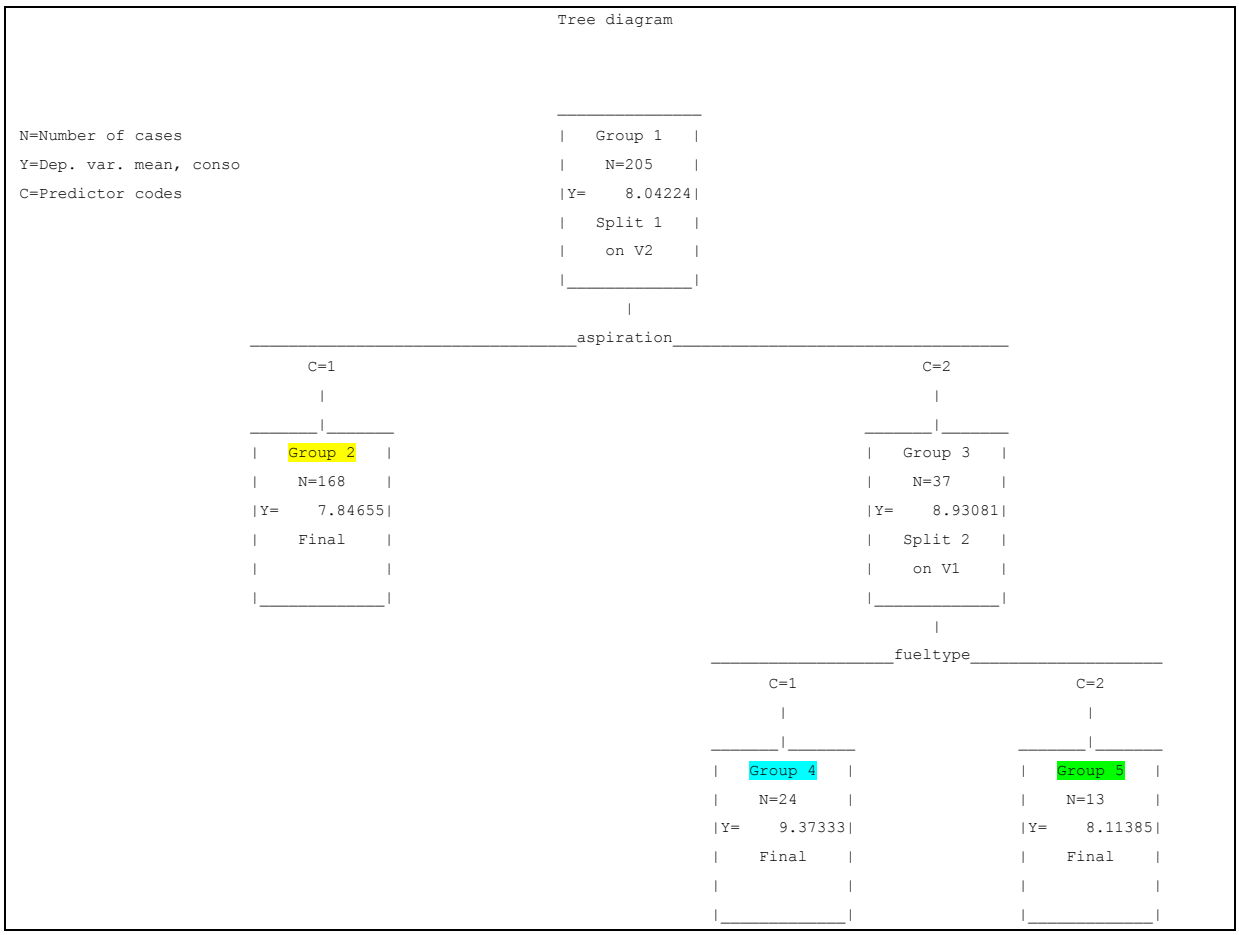

Voyons-en l'essentiel.

## 3.4 Lecture des résultats

### 3.4.1 Dessin de l'arbre

Les sommets sont numérotés (Group). Les effectifs (N) et les moyennes (Y) de la variable dépendante sont visibles. L'arbre comporte 3 feuilles (3 règles) :

- Si aspiration = std alors Moyenne(conso) = 7.84655 (N = 168)
- Si aspiration = turbo et fuel-type = gas alors Moyenne(conso) = 9.937333 (N = 24)
- Si aspiration = turbo et fule-type = diesel alors Moyenne(conso) = 8.11385 (N = 13)




Nous avons bien les mêmes informations lorsque nous croisons « aspiration » et « fuel-type » dans un tableau croisé où nous calculons les moyennes conditionnelles de « conso ».

| Moyenne d | Étiquet | | |
|---|---|---|---|
| Étiquett | 1 | 2 | Total |
| 1 | 7.9511 | 9.3733 | 8.1356 |
| 2 | 5.4414 | 8.1138 | 7.1785 |
| **Total** | **7.8465** | **8.9308** | **8.0422** |

aspiration

fuel-type

### 3.4.2 Caractérisation des groupes - Variation

Pour chaque feuille de l'arbre, WinIDAMS indique la moyenne, la variance, et la variation (section 56.1, pages 389 et 390).

```
Final group summary table

 Group    2         168 cases
     Mean(Y)=0.78465476E+01     Var(Y)=0.35703743E+01  Variation=0.59625250E+03

 Group    4          24 cases
     Mean(Y)=0.93733330E+01     Var(Y)=0.12849357E+01  Variation=0.29553522E+02

 Group    5          13 cases
     Mean(Y)=0.81138458E+01     Var(Y)=0.20273247E+01  Variation=0.24327898E+02
```

La variation correspond à la somme des carrés des écarts à la moyenne, soit

$$\text{Variation}(g) = (N_g - 1) * V_g$$

Où $N_g$ est l'effectif du groupe, $V_g$ sa variance estimée.

Pour le groupe n°2, nous avons par exemple :

$$(168 - 1) * 3.5704 = 596.25$$

### 3.4.3 Tableau d'analyse de variance

Le tableau d'analyse de variance se conçoit aisément dans le contexte d'une variable cible quantitative.

```
The partitioning ends with 3 final groups
The variation explained is 7.0%

 One-way analysis of final groups

         Source          Variation          DF

         Explained 0.49023743E+02           2
         Error     0.65013391E+03         202
         Total     0.69915765E+03         204
```

La proportion de variance expliquée est égale à

$$R^2 = 49.0237 / 699.1576 = 7\%$$

Ce n'est pas terrible. Le type de carburant et le mode d'alimentation expliquent peu la consommation. Ca paraît étrange quand même connaissant un peu les voitures.

### 3.4.4 Récapitulation des opérations de segmentation

Deux opérations de segmentation ont été initiées lors de la construction de l'arbre. Le logiciel propose un résumé des informations associées dans « Split Summary Table ».

```
Split summary table

 Group   1        205 cases
     Mean(Y)=0.80422440E+01     Var(Y)=0.34272432E+01  Variation=0.69915765E+03
     Split on V2      aspiration                        Var expl=0.35647308E+02
     Into group   2, codes  1
      and group   3, codes  2

Group   3         37 cases
     Mean(Y)=0.89308109E+01     Var(Y)=0.18682737E+01  Variation=0.67257851E+02
     Split on V1      fueltype                         Var expl=0.13376432E+02
     Into group   4, codes  1
      and group   5, codes  2
```

La première segmentation explique **5.10%** (35.6473 / 699.1577) de la variation de Y ; la seconde explique **1.91%** (13.3764 / 699.1577). Et **5.10% + 1.91% ≈ 7%** (aux arrondis près).

### 3.4.5 Impact global des variables

Tout comme pour l'arbre de décision, WinIDAMS propose un récapitulatif de l'impact des variables lors de chaque segmentation, dans le but d'identifier le rôle de celles qui n'apparaissent pas dans les règles de décision.

```
Per cent of total variation explained by best split for each group (*=Final groups)

                1       2*        3       4*       5*
 V1          2.36     0.00     1.91     0.00     0.00
 V2          5.10     0.00     0.00     0.00     0.00
```

Notre exemple est trop simple (2 variables explicatives seulement) pour que ce tableau soit réellement décisif. Mais dans un cas réel, avec un nombre plus important de variables candidates, ce tableau devient primordial.

### 3.4.6 Processus de construction

WinIDAMS propose une trace de toutes les opérations réalisées durant le processus de construction de l'arbre. Il s'agit d'une version détaillée du tableau de la section précédente (nous nous en tenons aux 3 premières segmentations).

```
#Un seul groupe (le 1er) peut être segmenté initialement
Split   1     candidate groups

     Group        N     Sum(WT)      Mean Y       Var Y       Variation
        1      205 0.20500E+03 0.80422E+01 0.34272E+01 0.69916E+03
```

```
#Tentative de segmentation du groupe 1
Attempt to split group   1       Var= 699.15765

Predictor V1      fueltype                  Rank 1  Type M
  Codes  1 2
  Best split after code  1  Var expl=0.16534172E+02

Predictor V2      aspiration                Rank 1  Type M
  Codes  1 2
  Best split after code  1  Var expl=0.35647308E+02

#V2 (aspiration) s'avère être la meilleure
Best split for group   1 on predictor V2      aspiration                Rank 1
     Var expl=0.35647308E+02

#Segmentation avec V2, création de 2 sous-groupes : les n°2 et 3
Split group   1 on V2      aspiration                  Var expl=0.35647308E+02
    Into group   2, codes  1
     and group   3, codes  2

#Etape 2 : 2 sommets peuvent être segmentés, le n°2 et n°3
Split   2     candidate groups

    Group       N    Sum(WT)      Mean Y      Var Y      Variation
       2      168 0.16800E+03 0.78465E+01 0.35704E+01 0.59625E+03
       3       37 0.37000E+02 0.89308E+01 0.18683E+01 0.67258E+02

#Tentative de segmentation du groupe n°2
Attempt to split group   2       Var= 596.25250

Predictor V1      fueltype                  Rank 1  Type M
  Codes  1 2
  No eligible split

Predictor V2      aspiration                Rank 1  Type M
  Codes  1
  No eligible split

#Aucune variable ne produit une segmentation éligible
No eligible split for group   2

#Il reste une segmentation à évaluer, celle du groupe n°3
Split   2     candidate groups

    Group       N    Sum(WT)      Mean Y      Var Y      Variation
       3       37 0.37000E+02 0.89308E+01 0.18683E+01 0.67258E+02
```

```
#Tentative de segmentation du groupe n°3
Attempt to split group   3       Var= 67.257851

Predictor V1      fueltype                  Rank 1   Type M
  Codes  1 2
  Best split after code  1  Var expl=0.13376432E+02

Predictor V2      aspiration                Rank 1   Type M
  Codes  2
  No eligible split

#V2 (aspiration) s'avère être la meilleure
Best split for group   3 on predictor V1      fueltype                  Rank 1
     Var expl=0.13376432E+02

#Segmentation avec V1, création de 2 sous-groupes : les n°4 et 5
Split group   3 on V1      fueltype                    Var expl=0.13376432E+02
    Into group   4, codes  1
     and group   5, codes  2

Etc.
```

Les informations proposées sont particulièrement complètes. Il est difficile de faire la fine bouche. Mais, d'un autre côté, les déchiffrer devient très rapidement ardu lorsque le nombre de variables et de sommets augmentent. On ne peut pas tout avoir.

### 3.5 Conclusion

Le principal atout de la procédure SEARCH est qu'elle inscrit dans un cadre unique cohérent le classement et la régression. Les sorties sont identiques, seule la définition de la variation change selon que l'on traite une cible qualitative ou quantitative. Pédagogiquement, la démarche est très intéressante.

## 4 Arbre de régression (II) – Arbre REGRESSION

### 4.1 Objet de la méthode

L'option ANALYSIS = REGRESSION de SEARCH vise à identifier les sous-groupes d'individus où les relations entre une variable DEPVAR et COVAR, sous la forme d'une régression linéaire, sont les plus pertinentes. Les groupes sont définis par un arbre de segmentation s'appuyant sur un ensemble de variables (VARS) discrètes. L'approche n'est pas commune. Si l'idée d'un test d'équivalence de régressions dans des sous-populations m'est familière – un test de Chow de rupture de structure répond exactement à cette spécification – je ne me rappelle

pas avoir rencontré dans un logiciel quelconque un algorithme visant à produire explicitement les sous-populations via un processus de partitionnement récursif. C'est l'étude de cette option qui m'a convaincu de ne pas intégrer SEARCH dans le document de présentation générique de WinIDAMS[13], et de lui consacrer un tutoriel spécifique.

## 4.2 Données

Nous reprenons les données « autos ». On cherche à expliquer la consommation des véhicules (DEPVAR = conso [V4]) à partir de leur cylindrée (COVAR = engine size [V3]). Les variables utilisées pour définir les sous-populations (pour construire les segmentations) sont « fuel-type » et « aspiration » [VARS = (V1-V2)].

## 4.3 Définition des traitements

Voici le fichier « setup » utilisé :

```
$RUN SEARCH
$FILES
PRINT   = regression_autos.lst
DICTIN  = autos.dic
DATAIN  = autos.dat
$SETUP
REGRESSION ANALYSIS FOR AUTOS DATASET
ANALYSIS=REGRESSION  -
DEPV=V4 -
COVAR=V3 -
MINC=10 -
PRINT=(FINAL,TREE,TRACE)
VARS=(V1-V2) TYPE=F
```

V1 et V2 étant binaires, le mode de regroupement (TYPE = F pour « free ») n'est pas opérant.

## 4.4 Lecture des résultats

### 4.4.1 Dessin de l'arbre

A l'instar des autres options, SEARCH produit un arbre décrivant la définition des groupes. Par rapport aux deux méthodes précédentes, les informations associées aux sommets sont un peu plus touffues.

[13] « Statistiques avec WinIDAMS », octobre 2014 ; http://tutoriels-data-mining.blogspot.fr/2014/10/statistiques-avec-winidams.html

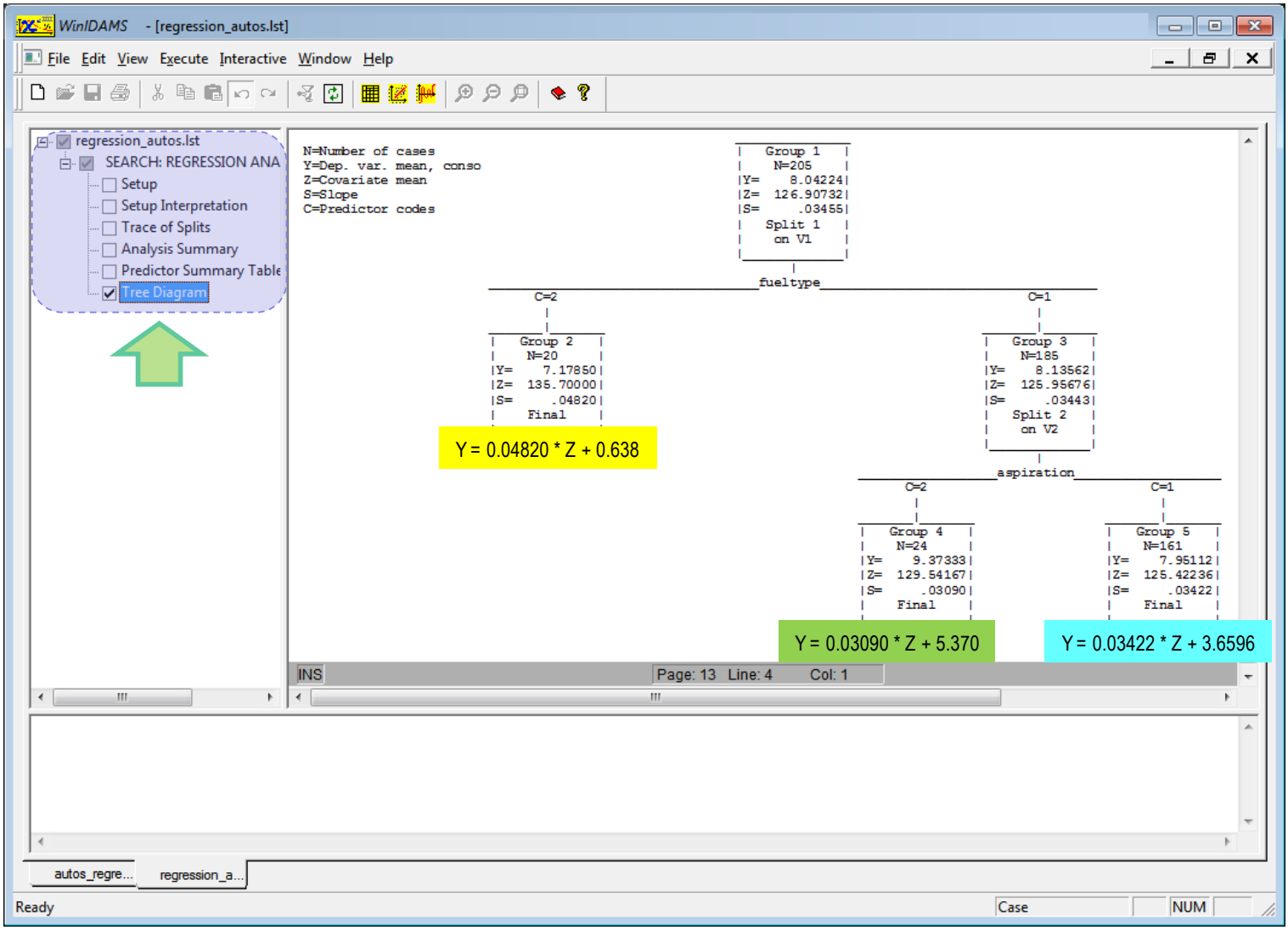



**Figure 3 - Arbre avec les régressions associées aux feuilles - Y : conso , Z : engine-size**

Nous disposons :

- de l'identifiant des groupes ;
- des effectifs (N) ;
- de la moyenne de la variable dépendante (Y) ;
- de la moyenne de la co-variable (Z) ;
- et de la pente (Slope) de la droite de régression associant Y à Z.

Nous observons que la première segmentation est réalisée à l'aide de « fuel-type » c.-à-d. la relation entre « conso » et « engine-size » est significativement différente selon le type de carburant (group 2 vs. group 3). Ensuite, chez les véhicules fonctionnant à l'essence (fuel-type = 1) (group 3), le mode d'alimentation (aspiration) permet d'opérer une distinction. A ce stade, on peut deviner la nature des segmentations en lisant la valeur de S (la pente de la

régression) dans les sommets : elles sont, selon le cas, basées sur une différenciation de la pente ou de l'origine. Nous essaierons de préciser ces idées ci-dessous.

### 4.4.2 Droites de régression – Modèle prédictif

Nous disposons des équations associées aux feuilles dans « Final group summary table » :

```
Final group summary table

Group    2          20 cases
    Mean(Y)=0.71785002E+01     Var(Y)=0.31229074E+01
    Mean(Z)=0.13570000E+03     Var(Z)=0.10313789E+04
      Slope=0.48199706E-01  Intercept=0.63779992E+00
       Corr=    0.87594E+00  Variation=0.13809118E+02

Group    4          24 cases
    Mean(Y)=0.93733330E+01     Var(Y)=0.12849357E+01
    Mean(Z)=0.12954167E+03     Var(Z)=0.53747644E+03
      Slope=0.30901793E-01  Intercept=0.53702636E+01
       Corr=    0.63201E+00  Variation=0.17748831E+02

Group    5          161 cases
    Mean(Y)=0.79511180E+01     Var(Y)=0.34467981E+01
    Mean(Z)=0.12542236E+03     Var(Z)=0.19983330E+04
      Slope=0.34216717E-01  Intercept=0.36595767E+01
       Corr=    0.82388E+00  Variation=0.17714919E+03
```

Nous avons pour chaque feuille de l'arbre:

- les moyennes (mean) et variance (Var) des deux variables Y (DEPVAR) et Z (COVAR) ;
- la pente (slope) et la constante (intercept) de l'équation de régression ;
- le coefficient de corrélation entre Y et Z (son carré est égal au coefficient de détermination de la régression) ;
- la variation (incertitude) associée au sommet.

A partir de ces informations, nous avons complété l'arbre en faisant figurer les équations de régressions sur les feuilles (Figure 3).

Nous pouvons également écrire le modèle sous la forme d'une base de règles permettant de prédire « conso » (Y) à partir de « engine-size »(Z) selon les valeurs de « fuel-type » et « aspiration » :

(Group 2) **Si** (fuel-type = diesel) **alors** Y = 0.04820 * Z + 0.638

(Group 4) **Si** (fuel-type = gas) **et** (aspiration = turbo) **alors** Y = 0.03090 * Z + 5.370

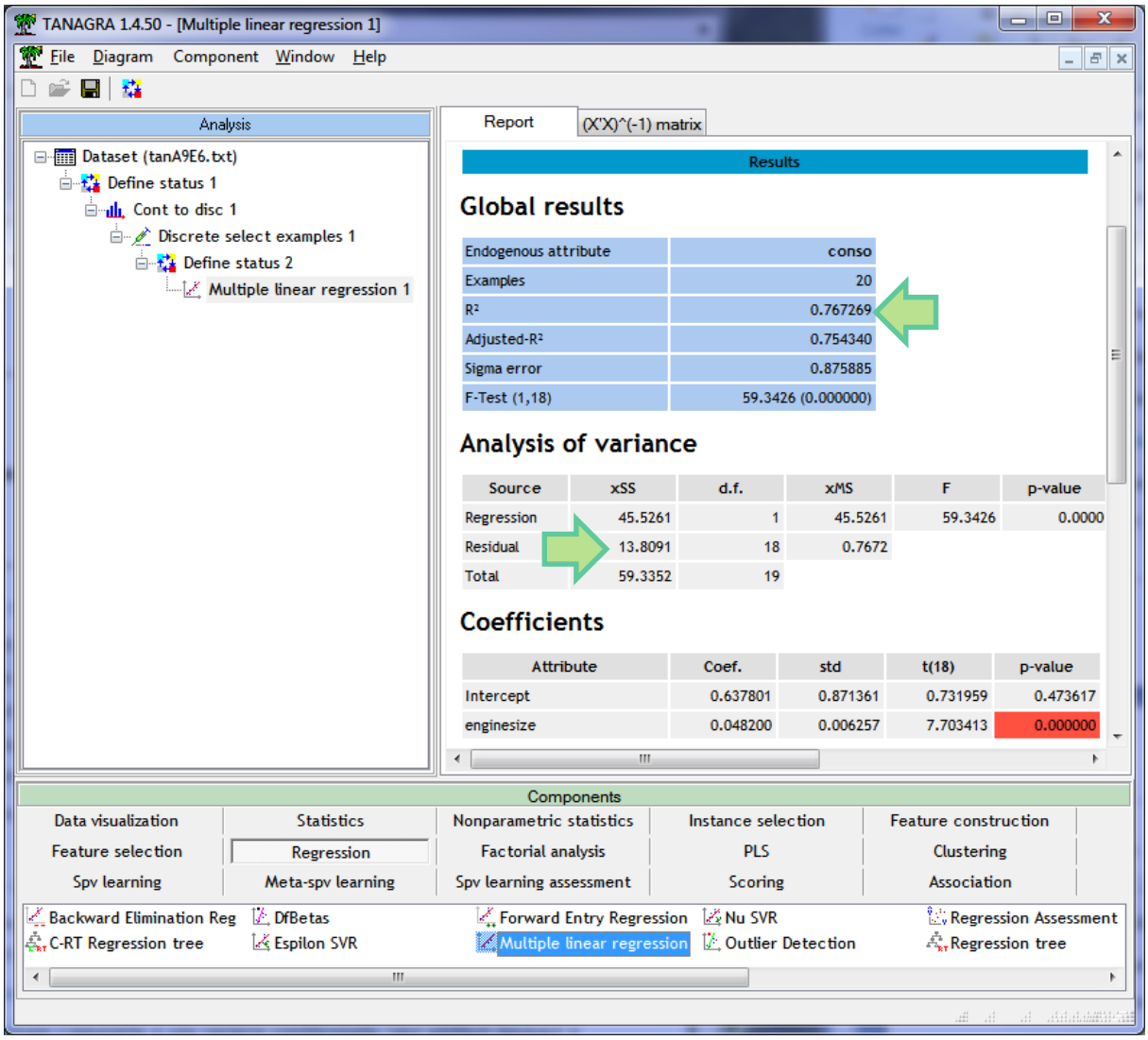
(Group 5) **Si** (fuel-type = gas) **et** (aspiration = std) **alors** Y = 0.03422 * Z + 3.6596

On peut imaginer – le concept est attrayant – qu'un pool de régressions optimisées pour des sous-populations est plus efficace en prédiction qu'une seule équation globale censée être valable pour l'ensemble de la population.

### 4.4.3 Variation dans les groupes

Le critère « variation » est central dans la procédure SEARCH. Il indique l'incertitude associée au sommet. L'objectif de la segmentation est de réduire cette incertitude. Dans le cas de REGRESSION, la variation correspond à la fraction d'information de Y non expliquée par Z dans la régression locale. Il s'agit ni plus ni moins que de la somme des carrés des résidus de la régression (section 56.2, page 391).

Pour le sommet n°2 (Group 2) constitué de 20 observations (fuel-type = diesel), nous obtenons à l'issue de la régression sous Tanagra :




La SCR = 13.8091, comme nous l'indique WinIDAMS; le coefficient de détermination de la régression est égal au carré de la corrélation entre Y et Z, soit $(0.8759^2) = 0.767269$.

### 4.4.4 Partition finale – Tableau d'analyse de variance

WinIDAMS propose le tableau d'analyse de variance. La variation totale correspond au SCR de l'équation de régression appliquée sur la totalité de l'échantillon. La variation résiduelle est égale à l'addition des variations sur les feuilles de l'arbre (SCR des régressions sur les sous-populations circonscrites par les feuilles). La variation expliquée est obtenue par différence.

```
The partitioning ends with 3 final groups

The variation explained is 24.6%

One-way analysis of final groups

        Source          Variation          DF
        Explained 0.68275726E+02            2
        Error     0.20870712E+03          202
        Total     0.27698285E+03          204
```

La proportion de variation expliquée par l'arbre est

$$68.275726 / 276.98285 = \mathbf{24.6}\%$$

### 4.4.5 Processus de construction

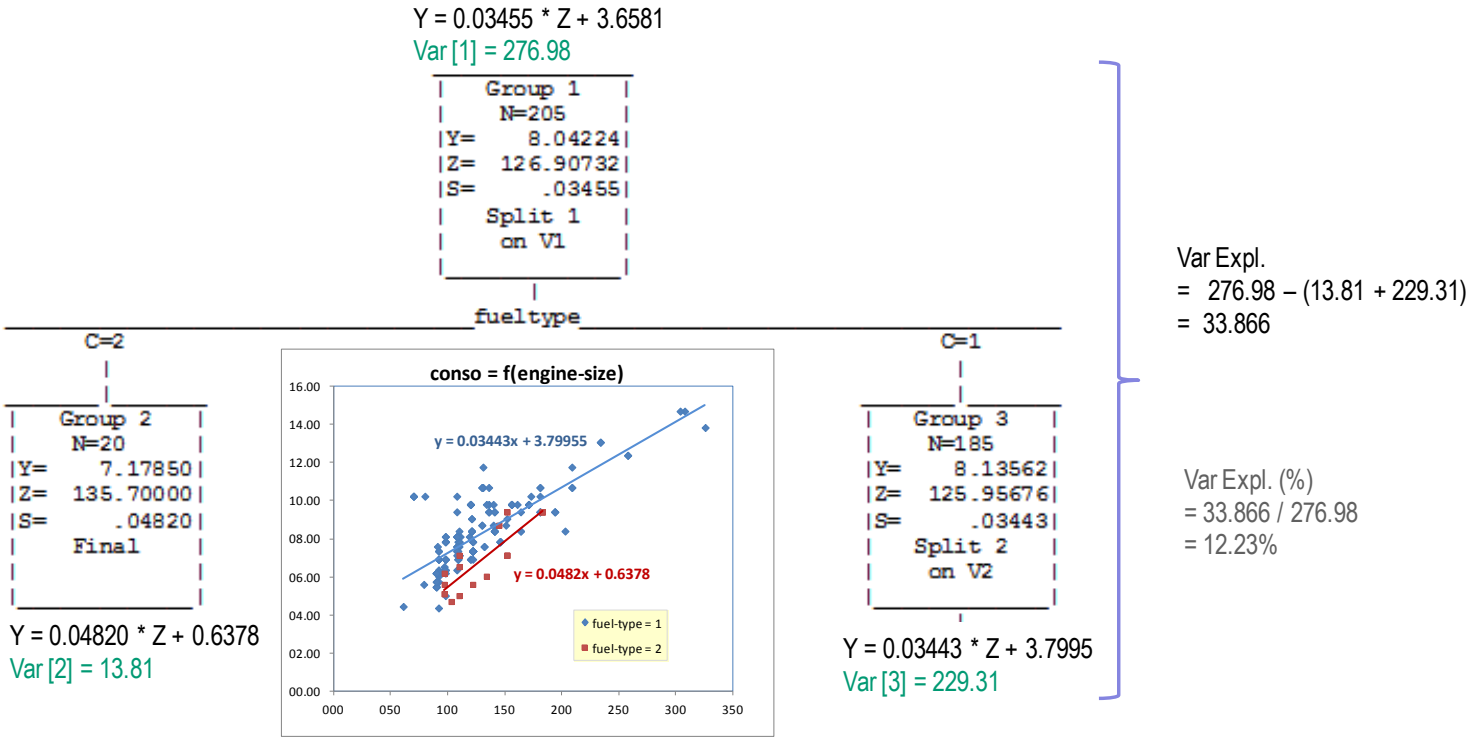
WinIDAMS fournit tous les éléments de compréhension du processus de partitionnement. Nous retraçons graphiquement les régressions à opposer pour chaque sommet à segmenter.

**Traitement du groupe 1**. La racine de l'arbre est segmentée à l'aide de la variable « fuel-type ». La variation expliquée par le découpage est égal à **33.87**. Ramenée à la variation initiale, elle est égale à **12.23%**.

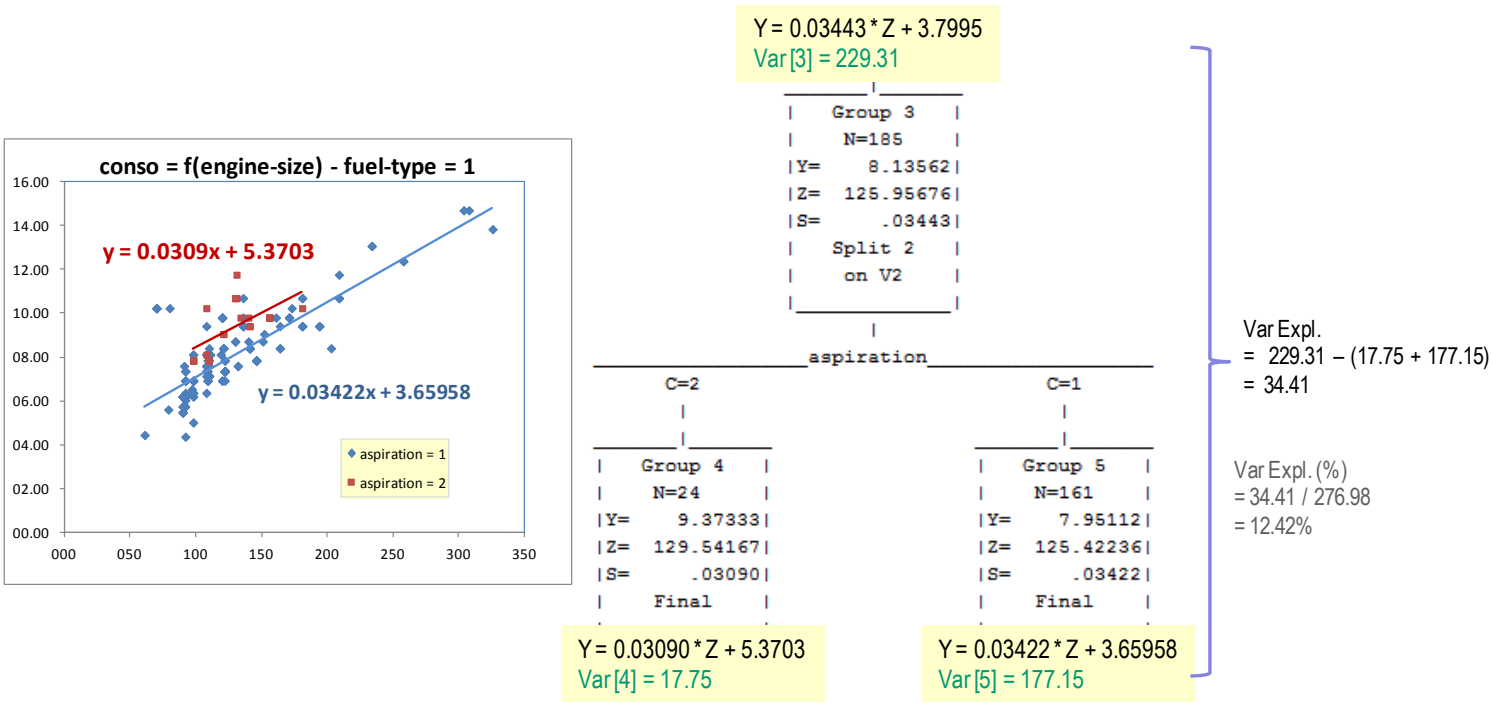
```
Split summary table

Group   1        205 cases
     Mean(Y)=0.80422440E+01     Var(Y)=0.34272435E+01
     Mean(Z)=0.12690732E+03     Var(Z)=0.17341140E+04
       Slope=0.34545563E-01  Intercept=0.36581593E+01
        Corr=0.77706724E+00  Variation=0.27698285E+03
     Split on V1      fueltype                        Var expl=0.33866150E+02
     Into group   2, codes  2
      and group   3, codes  1
```

L'équation de régression sur la racine s'écrit : Y = 0.03455 * Z + 3.6581.

Après la partition, nous avons 2 modèles distincts. En visualisant les sous-populations associées dans le plan (Z, Y), nous constatons que la partition repose avant tout sur la différenciation des pentes de régression.




**Traitement du groupe 3**. Le sommet n°3 est segmenté à l'aide de la variable « aspiration » avec une variation expliquée de 34.4096.




La différence repose sur un écart entre les constantes (entre les origines) des régressions cette fois-ci. Les deux droites sont approximativement parallèles.

### 4.4.6 Impact global des variables

Pour situer le rôle des variables qui n'apparaissent pas explicitement dans l'arbre, WinIDAMS fournit un résumé variations expliquées sur l'ensemble des sommets.

```
Per cent of total variation explained by best split for each group (*=Final groups)

                1       2*        3       4*       5*
 V1         12.23     0.00      0.00     0.00     0.00
 V2          6.50     0.00     12.42     0.00     0.00
```

## 4.5 Conclusion

Une fois comprise la finalité de la méthode, l'option REGRESSION s'inscrit parfaitement dans la procédure SEARCH, avec des structures de sorties alignées sur MEANS et CHI. L'enjeu finalement réside dans la définition de la « variation ». Dans le cadre de la régression, traduire l'idée d'incertitude ou de dispersion non-expliquée à l'aide de la somme des carrés des résidus de la régression est parfaitement cohérente avec le processus de segmentation.

On perçoit aisément les extensions possibles de l'approche. On pourrait passer à une régression multiple en adoptant plusieurs variables explicatives, traiter une variable dépendante qualitative binaire en s'appuyant sur une régression logistique. La déviance ferait office de mesure de variation dans ce cas. Les pistes sont nombreuses. On en entend parler d'ailleurs ici ou là dans les articles scientifiques. Mais ces techniques ne sont jamais programmées dans des logiciels accessibles à tous. Personne ne peut les mettre en œuvre dans des études réelles. Ce qui annihile de facto leur portée. Un des grands mérites de WinIDAMS est de proposer un outil qui marche, et qui est utilisable sur nos propres données.

# 5 Conclusion

La cohérence scientifique est certainement l'aspect le plus séduisant de la procédure SEARCH. Le cadre et la démarche sont toujours les mêmes. Seule la définition de la variation, mesure d'incertitude associée aux nœuds, est différente selon le problème que l'on traite.

Le paramétrage de l'outil reste relativement simple. Les effectifs des nœuds permettent de contrôler la taille de l'arbre. Non exploité dans ce tutoriel, il est également possible de définir un gain minimal pour la segmentation (option EXPL, section 36.7, page 264).

S'agissant des arbres, l'expérience montre que la possibilité pour l'utilisateur d'intervenir lors de la construction du modèle pour le guider vers des solutions en adéquation avec les connaissances du domaine est un atout fort. WinIDAMS propose des sorties au format texte. Procéder à une analyse interactive où l'on pourrait intervenir sur les nœuds pour élaguer ou

initier ses propres découpages n'est pas possible. En revanche, et c'est une fonctionnalité originale par rapport aux outils du même genre (avec des sorties non interactives), il est possible avec le paramètre GNUM de spécifier explicitement les variables et regroupements à exploiter sur les nœuds (section 36.7, page 266, « Predefined split specifications »). Manifestement, une réflexion a été menée pour intégrer la possibilité de guider l'induction. La solution proposée est souple et fonctionnelle.

Enfin, les sorties particulièrement touffues de la procédure SEARCH peuvent déconcerter au premier abord. J'ai essayé dans ce tutoriel de les démêler en mettant en exergue les résultats importants. On se rend compte que les éléments fournis par WinIDAMS permettent de retracer dans le détail le processus de modélisation. Ils peuvent aussi nous fournir les arguments pour guider la construction de l'arbre à l'aide de l'option GNUM.

Autant j'ai été réservé concernant WinIDAMS dans mon précédent tutoriel de présentation générale du logiciel[14] (logiciel pas très « user friendly », sorties non alignées sur les standards du domaine…), autant je suis enthousiaste concernant la procédure SEARCH.

[14] http://tutoriels-data-mining.blogspot.fr/2014/10/statistiques-avec-winidams.html