SPAD•S

VERSION 5.0

Procédures de segmentation

CISIA-CERESTA - 261 rue de Paris - 93556 MONTREUIL Cedex

tel : +33 (0)1 55 82 15 15 - Fax : +33 (0)1 43 63 21 00

e-mail : cisia@fr.inter.net - Web : http://www.cisia.com

SPAD•S

Option Segmentation du logiciel SPAD

Segmentation
par
Arbre de décision binaire
Discrimination et Régression

MANUEL DE REFERENCE

SPAD•S®
Segmentation par Arbre de Décision Binaire
Discrimination et Régression.

A. GUEGUEN, J-P. NAKACHE, J. NICOLAU-MOLINA
INSERM, Unité 88 Recherche Clinique et Biostatistique

avec la collaboration de
M-P. Bayol, A. Morineau, P. Pleuvret

ISBN 2-906711-20-9

Centre **I**nternational de **S**tatistique et d'**I**nformatique **A**ppliquées
261 rue de Paris, 93556 Montreuil Cedex (France)
Tel : 01 55 82 15 15 – Fax : 01 43 63 21 00
e-mail : cisia@fr.inter.net – Web : http://www.cisia.com/

Avant Propos

“Segmentation”, “discrimination”, “classement”, “classification” ou “classification supervisée” ... — le vocabulaire est riche pour désigner, suivant le domaine d’application, des opérations qui sont souvent proches sinon identiques.

Le logiciel, dont le fonctionnement est détaillé dans ce manuel, résout le problème de la discrimination et de la régression en terme d’**arbre binaire**. La voie a été ouverte par Morgan et Sonquist (1963) et Morgan et Messenger (1973) avec la méthode dite AID (Automatic Interaction Detection). De nombreuses contributions ont suivi, jusqu’à la parution de l’ouvrage de Breiman, Friedman, Olshen et Stone en 1984. Leur méthode repérée par l’acronyme CART (pour “Classification And Regression Tree”) diffère de AID par le mode de construction de l’arbre et la technique *d’élagage* conduisant à un sous-arbre exploitable et en quelque sorte “optimal”.

La segmentation par arbre de décision binaire présente des avantages importants. Le premier est sans doute la lisibilité des règles d’affectation: l’interprétation des résultats est directe et intuitive. Par ailleurs la technique est non-paramétrique (qualité à juste titre prisée par l’utilisateur) et peu contrainte par la nature des données. On peut en effet mettre en même temps, parmi les variables “explicatives”, des variables continues, ordinales et nominales. Noter que la technique fournit d’office la sélection des variables à utiliser. Enfin c’est le même principe, la même méthode, le même algorithme qui sont mis en œuvre pour analyser une variable nominale (analyse discriminante) et une variable continue (régression multiple).

Cependant les règles d’affectation pourront paraître parfois “abruptes” et trop vite sensibles à de légères perturbations des données. Il apparaîtra parfois difficile de décider quel est l’arbre “optimal”. Certains utilisateurs pourront aussi regretter l’absence d’une fonction globale mettant en jeu l’ensemble des variables (fonction linéaire discriminante ou équation de régression), perdant ainsi la représentation géométrique sous forme de configurations de points dans l’espace.

Les procédures constituant le logiciel SPAD•S ont été programmées à partir des éléments théoriques publiés dans l'ouvrage de Breiman *et al.*, par J. Nicolau-Molina sous la direction de A. Gueguen et J-P. Nakache dans l'équipe Recherche Clinique et Biostatistique de l'Unité INSERM U88. On peut se reporter à l'article de A. Gueguen et J-P. Nakache publié dans la *Revue de Statistique Appliquée* pour une présentation pratique de cette méthode illustrée à partir d'un exemple médical et à l'ouvrage "*Analyse Discriminante sur Variables Qualitatives*" paru chez Poytechnica (1994, 270 pages).

Références générales

L. BREIMAN, J.H. FRIEDMAN, R.A. OLSHEN and C.J. STONE (1984).- *Classification and Regression Trees*. Wadwsorth International Group.

G. CELEUX (1990).- *Analyse Discriminante sur Variables Continues.* INRIA.

G. CELEUX, J.P. NAKACHE (1994).- *Analyse Discriminante sur Variables Qualitatives*. Polytechnica, Paris.

A. CIAMPI, C.H. CHANG, S. HOGG, S. MCKINNEY (1987).- *Recursive Partition : A versatile method for exploratory data analysis in biostatistics.* Proccedings from Joshi Festschrift, G. Humprey Ed. pp.23-50. Amsterdam : North Holland.

A. GUEGUEN, J.P. NAKACHE (1988).- *Méthode de discrimination basée sur la construction d'un arbre de décision binaire*: Revue de Statistique Appliquée, XXXVI (1), pp.19 - 38.

G.V. KASS (1980).- *An Exploratory Technique for Investigating Large Quantities of Categorical Data*. Applied Statistics, pp.29, 119-127.

J.N. MORGAN and J.A. SONQUIST (1963) - *Problems in the Analysis of Survey Data and a Proposal*. J. Amer. Statist. Assoc., vol.58, pp.415-435.

J.N. MORGAN and R.C. MESSENGER (1973) - *THAID: a Sequential Search Program for the Analysis of Nominal Scale Dependent Variables.* Institute for Social Research, Univ. Michigan.

J. NICOLAU MOLINA (1989).- *Classification pronostique des cancers du sein polymestastasés traités par chimiothérapie.* Rapport DEA Biomathématiques, Université de Paris 7.

J.A. SONQUIST and J.N. MORGAN (1964) - *The Detection of Interaction Effects*. Ann Arbor : Institute for Social Research, University of Michigan.

Sommaire

Partie 1 :

Notices des Procédures
et
description des méthodes

Procédure COLIG :
Transposition du tableau de données

1. Présentation

1.1 Objet

Cette procédure est une étape de gestion préliminaire et obligatoire avant toute procédure de construction d'un arbre de décision binaire.

La procédure transpose et ordonne le tableau des données créé par SELEC, afin d'assurer un traitement rapide par les procédures suivantes.
La procédure crée un nouveau fichier de type NSEG.

1.2 Editions

Il n'y a pas d'autre édition que les caractéristiques du fichier NSEG créé par la procédure.

1.3 Paramètres

La procédure ne requiert aucun paramètre.

2. Instructions de commande

(1) PROC COLIG Transposition et rangement des données

(2) *Titre de la procédure*

3. Exemple de commande

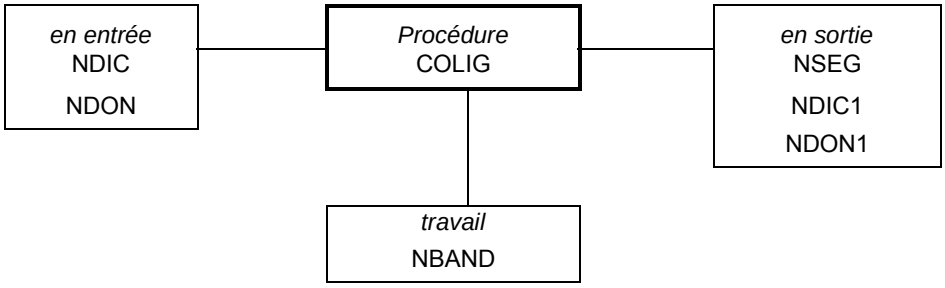
```
----+----1----+----2----+----3----+----4----+----5----+----6---
LISTP = OUI
PROC COLIG
transposition Préliminaire

----+----1----+----2----+----3----+----4----+----5----+----6---
```

4. Fichiers nécessaires à l'exécution

- en lecture NDIC (fichier dictionnaire utile)
  NDON (fichier des données utiles)

- en écriture NSEG (fichier des données transposées)
  NDIC1 (fichier dictionnaire utile apuré)
  NDON1 (fichier des données utiles apuré)

- de travail NBAND (fichier non formatté)



Les fichiers NDIC1 et NDON1 sont les nouveaux fichiers utiles (dictionnaire et données) apurés des modalités vides pour les variables nominales.

Procédure DISAR :
Discrimination par arbre de segmentation

1. Présentation

1.1 Objet

Cette procédure effectue la construction d'un arbre de décision binaire complet pour la discrimination à k groupes caractérisés par les modalités d'une variable nominale. Elle prend en compte des variables explicatives continues, nominales et ordinales et permet de tenir compte des coûts de mauvaise classification et des probabilités a priori d'appartenance aux groupes.

La méthode fournit, à partir de l'arbre complet, la séquence des sous-arbres obtenue en utilisant une procédure d'élagage basée sur la suppression successive des branches les moins informatives en terme de discrimination entre les classes. Dans cette séquence, elle sélectionne un sous-arbre "optimal" à l'aide de l'échantillon de base en se basant sur l'estimation du taux d'erreur théorique de classement.

Les variables explicatives ne doivent présenter aucune donnée manquante. Les données déclarées "manquantes" (code 0) dans la variable nominale à discriminer repèrent en réalité les individus dits "anonymes".

L'étape DISAR doit être précédée de la procédure de gestion COLIG utilisée pour transposer le tableau de données. La procédure utilise en entrée les fichiers de type NDIC et NSEG.

1.2 Editions

La procédure permet de tracer l'arbre si le nombre de niveaux est inférieur ou égal à 13. Quel que soit le nombre de niveaux, elle donne une description de l'arbre: description des divisions et taille des segments. Le programme fournit en option la séquence d'élagage, et pour tout arbre de cette séquence, l'ensemble de ses segments terminaux.

Enfin le programme imprime les informations qui permettent de déterminer le sous-arbre optimal : à chaque arbre de la séquence est associé l'indice de risque (coût relatif) de l'échantillon-test et de l'échantillon de base.

1.3 Paramètres

Les paramètres de la procédure se divisent en trois catégories:

- **Les paramètres de définition de l'analyse** : la variable à expliquer (VAR); la présence de variables ordinales (LORD); le type de la matrice des coûts (LCOUT); le type du vecteur des probabilités a priori (LPROB); le pourcentage d'individus pour constituer l'échantillon-test (PRCT); ou le numéro de la variable servant à définir l'échantillon-test (IVTES).

- **Les paramètres de fonctionnement** permettent de fixer les caractéristiques de l'arbre de décision : effectif minimum requis pour diviser un segment (NXIND); nombre maximum de niveaux de l'arbre (NXNIV); nombre maximum de segments de l'arbre (NXNOD) ; nombre de groupes pour la validation croisée (NBGVC).

- **Les paramètres d'édition** : nombre de divisions "équi-réductrices" (ou concurrentes) (NRED); nombre de divisions "équi-divisantes" (ou suppléantes) (NDIV); la description complète de l'arbre (LDESC); le dessin de l'arbre complet (LDESS); la description de la séquence d'élagage (LSEQU); l'édition de l'information détaillée de la validation croisée (LDETA).

1.4 Commandes par listes

Si on a déclaré soit des variables ordinales (LORD = OUI), soit une matrice des coûts (LCOUT = LIST), soit un vecteur de probabilités a priori (LPROB = LIST), les listes correspondantes devront suivre **immédiatement**. Elles seront écrites avec le format SPAD•N, et ne contiendront ni ligne vide ni commentaire (il s'agit de **données** et non d'instructions).

2. Instructions de commande

Les valeurs par défaut sont indiquées entre parenthèses, à la suite du nom du paramètre.

(1) PROC DISAR Analyse discriminante par arbre de segmentation

(2) *titre donné à l'analyse discriminante*

(3) VAR numéro de la variable à expliquer (pas de valeur par défaut).

PRCT (33.0) pourcentage d'individus dans l'échantillon-test.

IVTES (0) numéro de la variable servant à définir un échantillon-test

LORD (0 ou NON) présence d'une liste de variables ordinales.
- 0 ou NON : pas de variables ordinales.
- 1 ou OUI : il y a une liste de variables ordinales.

NXIND (5) nombre minimum d'individus dans un segment divisible.

NXNIV (13) nombre maximum de niveaux de l'arbre.

NXNOD (200) nombre maximum de segments.

LCOUT (1 ou UNIF) définition de la matrice des coûts.
- 1 ou UNIF : matrice de coûts unitaires.
- 2 ou LIST : la matrice est fournie par l'utilisateur.

LPROB (3 ou PROP) définition de la matrice des probabilités.
1 ou UNIF : probabilités égales.
2 ou LIST : probabilités données par l'utilisateur.
3 ou PROP : probabilités proportionnelles aux effectifs des classes.

NRED (0) nombre de divisions équi-réductrices (ou concurrentes) à éditer.

NDIV (0) nombre de divisions équi-divisantes (ou suppléantes) à éditer.

LDESC (0 ou NON) description de l'arbre complet.
1 ou OUI : édition de l'arbre.
0 ou NON : pas d'édition.

LDESS (0 ou NON) dessin de l'arbre complet

LSEQU (0 ou NON) description de la séquence d'élagage.

NBGVC (10) nombre de groupes pour la validation croisée.

(4) Liste des variables ordinales (si LORD = OUI)

(5) Matrice des coûts (si LCOUT = LIST)

(6) Vecteur des probabilités a priori (si LPROB = LIST)

3. Présentation détaillée des paramètres

VAR **numéro de la variable à expliquer**

- *valeurs possibles* : entières de 1 à NQEXA
- ***pas de valeur par défaut***

La variable à expliquer est une variable nominale avec k ≥ 2 modalités. Une valeur manquante pour cette variable (codée 0) correspond à un individu **anonyme**. Les individus anonymes ne sont pas utilisés dans l'analyse : ils sont affectés à l'un des groupes, une fois déterminé le sous-arbre optimal.

Remarque : Pour avoir des individus anonymes, il faut avoir codé LZERO = NOREC dans la procédure SELEC afin que les valeurs manquantes conservent la valeur zéro.

PRCT **pourcentage d'individus dans l'échantillon - test**

- *valeurs possibles* : réelles non nulles
- *valeur par défaut* : 33.0

PRCT représente le pourcentage d'individus tirés au hasard dans l'échantillon total pour constituer un échantillon-test. Le reste des individus, appelé échantillon de base, sert à construire l'arbre complet et la séquence des sous-arbres.

Dans la méthode de segmentation, l'échantillon-test sert à sélectionner un sous-arbre "optimal".
Si PRCT = 0, il n'y a pas d'élagage du grand arbre (qui est construit).

Remarques :
1) Si l'échantillon total est de petite taille et que l'on ne veut pas par conséquent le diviser en échantillon de base et échantillon-test, on peut utiliser la validation croisée.

2) L'échantillon-test est constitué de PRCT % d'individus tirés au hasard dans chacun des groupes à discriminer.

3) Si on désire faire l'analyse avec deux échantillons-tests distincts, il est indispensable de changer le paramètre général de tirage au hasard NAPEL au début de la procédure DISAR.

Exemple :

- *1ER ECHANTILLON-TEST* :

```
NAPEL = 0
PROC DISAR
ANALYSE DISCRIMINANTE PAR SEGMENTATION
VAR = 1
```

- *2E ECHANTILLON-TEST* :

```
NAPEL = 100
PROC DISAR
ANALYSE DISCRIMINANTE PAR SEGMENTATION
VAR = 1
```

IVTES **Identification de la variable servant à définir un échantillon-test**

- *valeurs possibles* : entiers de 1 à NQEXA
- *valeur par défaut* : 0

Cette variable servant à définir l'échantillon-test doit prendre deux valeurs:
- **1** pour repérer les individus appartenant à **l'échantillon-test**
- **2** pour repérer les individus appartenant à **l'échantillon de base**

Au sens de SPAD, c'est donc une variable nominale à 2 modalités.
Si IVTES = 0, on n'utilise pas de variable pour définir l'échantillon-test. Dans ce cas, l'échantillon-test est défini par PRCT :
* Si PRCT = 0, on ne fera pas d'élagage de l'arbre.
* Si PRCT > 0, l'échantillon-test sera défini par tirage aléatoire.

Si IVTES > 0, la valeur IVTES indique le numéro de la variable servant à définir l'échantillon-test.

Remarque :
La définition d'un échantillon-test est obligatoire, si on n'utilise pas la validation croisée. Cet échantillon peut être défini par le paramètre IVTES, ou par le paramètre PRCT. En cas de double définition, IVTES est prioritaire.

LORD **présence d'une liste de variables qualitatives à considérer comme ordinales**

- *valeurs possibles* : 0 ou NON pas de variables ordinales
  1 ou OUI présence d'une liste de variables ordinales
- *valeur par défaut* : 0 ou NON

Dans la méthode de segmentation, les variables ordinales jouent un rôle particulier. Si LORD = 1 ou OUI, la liste des variables ordinales est introduite par le mot clé ORDRE après la liste des paramètres généraux. Ces variables doivent avoir été sélectionnées comme variables nominales dans la procédure SELEC qui a précédé.

NXIND **nombre minimum d'individus dans un segment « divisible »**

- *valeurs possibles* : entières positives
- *valeur par défaut* : 5

Le choix NXIND = 1 conduit à des arbres de grande taille, pouvant avoir autant de segments terminaux que d'individus.
La valeur par défaut convient en général. NXIND est un des paramètres qui arrête la division des segments.
Remarque: le segment concerné pouvant être divisé, il sera normal de trouver des segments terminaux où l'effectif est inférieur à NXIND.

NXNIV **nombre maximum de niveaux de l'arbre**

- *valeurs possibles* : entières de 1 à 32
- *valeur par défaut* : 13

Le nombre de niveaux d'un arbre se compte de la façon suivante : le premier segment constitue le niveau 0; ses deux descendants immédiats sont au niveau 1, etc. La valeur du paramètre NXNIV indique le nombre maximum de niveaux que peut atteindre "l'arbre maximum" ou "grand arbre". NXNIV est un des paramètres qui arrête la division des segments.

Remarque : Si un arbre a plus de 13 niveaux, il est construit mais il n'est pas dessiné.

NXNOD **nombre maximum de segments de l'arbre**

- *valeurs possibles* : entières positives
- *valeur par défaut* : 200

La construction de l'arbre maximum s'arrête dès que le nombre de segments est égal à NXNOD. NXNOD est un paramètre de contrôle qui arrête la division des segments.

Attention : Si le nombre total de segments édités dans la description de l'arbre complet avant élagage est voisin de 200 (valeur par défaut), l'arbre a beaucoup de chances de ne pas être complètement construit. Dans ce cas il faut relancer la procédure en augmentant la valeur de NXNOD.

LCOUT **matrice des coûts**

- *valeurs possibles* : 1 ou UNIF matrice des coûts unitaires
  2 ou LIST matrice des coûts donnée par l'utilisateur
- *valeur par défaut* : 1 ou UNIF

Si LCOUT = 2 ou LIST, la matrice des coûts associée aux catégories à discriminer est introduite par le mot clé COUT après la liste des paramètres généraux.

LPROB **définition de la matrice des probabilités associées**

- *valeurs possibles* : 1 ou UNIF probabilités égales
  2 ou LIST probabilités données par l'utilisateur
  3 ou PROP probabilités proportionnelles aux effectifs des classes
- *valeur par défaut* : 3 ou PROP

Si LPROB = 2 ou LIST, la matrice des probabilités associée aux catégories à discriminer est introduite par le mot clé PROB après les paramètres généraux.

NRED **nombre de divisions équi-réductrices (ou concurrentes) à éditer**

- *valeurs possibles* : entières positives ou nulles
- *valeur par défaut* : 0

La meilleure division est celle qui assure la plus grande réduction d'impureté en passant d'un segment parent à ses segments descendants.
Par extension, on appelle "equi-réductrices" ou "concurrentes" les divisions qui assurent les plus fortes réductions d'impureté. Par exemple si NRED = 2 on obtient les deux meilleures divisions après la division fournie (qui est en réalité la meilleure).

Ce paramètre est utilisé au moment de l'édition de la description des divisions. Au maximum NRED divisions équi-réductrices sont éditées. En pratique, deux à cinq divisions équi-réductrices sont en général suffisantes. (Voir définition détaillée dans paragraphe 3 : "autres caractéristiques de DISAR").

NDIV **nombre de divisions équi-divisantes (ou suppléantes) à éditer**

- *valeurs possibles* : entières positives ou nulles
- *valeur par défaut* : 0

Par extension, on appelle "équi-divisantes" les divisions qui fournissent les répartitions les plus proches de la meilleure division.
Si NDIV = 2 on obtient les deux divisions qui assurent les répartitions les plus proches de la répartition considérée.
Ce paramètre est utilisé au moment de l'édition de la description des divisions. Au maximum NDIV divisions équi-divisantes sont éditées. En pratique, deux à cinq divisions équi-divisantes sont en général suffisantes. (Voir définition détaillée dans paragraphe 3 : "autres caractéristiques de DISAR").

Le code « c » à côté des variables équi-divisantes indique qu'il faut inverser les règles d'affectation.

LDESC **description de l'arbre complet**

- *valeurs possibles* : 0 ou NON pas de description de l'arbre
  1 ou OUI description de l'arbre
- *valeur par défaut* : 0 ou NON

Si LDESC = OUI, on obtient l'édition de la description de l'arbre complet. Cette édition peut être volumineuse et l'arbre difficile à lire. L'étape DISEL (choix de l'arbre élagué) permet l'édition du sous-arbre choisi.

LDESS **dessin de l'arbre complet**

- *valeurs possibles* : 0 ou NON pas de dessin de l'arbre
  1 ou OUI dessin de l'arbre
- *valeur par défaut* : 0 ou NON

NB : Même si on le demande, l'arbre n'est pas dessiné s'il possède plus de 13 niveaux.

LSEQU **description de la séquence d'élagage**

- *valeurs possibles* : 0 ou NON pas de description
  1 ou OUI description de la séquence d'élagage.
- *valeur par défaut* : 0 ou NON

Remarques générales

• Si le problème met en jeu plusieurs variables nominales ayant un grand nombre de modalités, le nombre de divisions dans la procédure est très important.
Par exemple, pour une variable nominale à 10 modalités, le nombre de divisions sera égal à $2^{10-1} - 1 = 511$; ce nombre est égal à 16383 si la variable nominale possède 15 modalités. Le programme utiliserait donc dans ce cas un volume important de mémoire. En cas de nécessité, on procédera à des regroupements de modalités.

• Lorsque l'on a des variables ordinales avec des réponses manquantes, et que l'on a codé REC dans SELEC, les réponses manquantes ont la valeur n + 1, et seront donc du côté des valeurs les plus fortes. Il est recommandé de se débarrasser des valeurs manquantes.

NBGVC **nombre de sous-groupes pour la validation croisée**

- *valeurs possibles* : 2 à NIEXA
- *valeur par défaut* : 10

La méthode de validation croisée est utile dans le cas de petits échantillons. Elle permet de prendre en compte tous les sujets de l'échantillon à la fois pour construire l'arbre (échantillon de base) et pour le tester (échantillon-test).

L'échantillon total est divisé de manière aléatoire en NBGVC sous-ensembles disjoints de même taille L_1, L_2, ... , L_k. Ainsi à chaque échantillon L_i (correspondant à un échantillon-test) est associé l'échantillon L^i complémentaire de L_i dans l'échantillon total (correspondant à un échantillon de base).

4. Listes des sélections

Après la liste de paramètres, il faut insérer la liste des variables ordinales (si LORD = OUI), puis la matrice des coûts (si LCOUT = LIST) et enfin le vecteur des probabilités a priori (si LPROB = LIST).

La syntaxe à employer suit le schéma suivant:

mot-clé *liste de valeurs*

• **mot-clé** est l'un des trois mots-clés suivants:

- ORDRE pour la liste des variables nominales à considérer comme ordinales
- COUT pour la matrice des coûts d'ordre égal au nombre de groupes à déterminer
  Remarque : Pour deux groupes par exemple on écrira :
  soit C_{11} C_{12} >
  C_{21} C_{22}
  soit C_{11} C_{12} C_{21} C_{22}
- PROB pour le vecteur de probabilités a priori

• *Liste de valeurs* est une liste de nombres. Le séparateur entre les nombres peut être le blanc ou la virgule.

Une telle liste est une instruction SPAD: on peut donc utiliser le caractère de continuation ">" ainsi que le symbole de suite "--" pour la liste des variables ordinales.

ORDRE: Tous les numéros présents dans la liste des variables ordinales doivent être entiers et doivent correspondre à des variables nominales sélectionnées par SELEC (nombres positifs et inférieurs ou égaux au nombre total de variables (NQEXA)).

COUT: Tous les nombres présents dans la matrice des coûts doivent être positifs ou nuls. La matrice des coûts constitue une matrice carrée dont le nombre de lignes et de colonnes est égal au nombre de classes de la variable expliquée. Le coût de mauvaise classification d'une classe dans elle-même est 0, donc la diagonale principale doit toujours être composée de 0. Après le mot-clé COUT, on écrit la liste des valeurs des coûts en mettant bout-à-bout les lignes de cette matrice des coûts. On peut faire apparaître l'image de la matrice en utilisant le symbole de continuation à la fin des lignes (voir l'exemple).

PROB : On écrit les probabilités a priori choisies par l'utilisateur pour chacun des groupes.

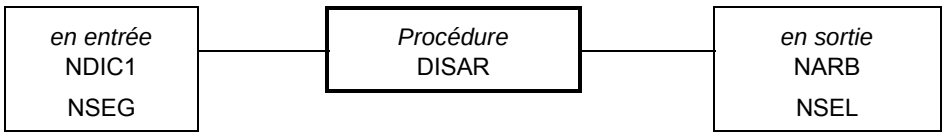
5. Exemple

```
----+----1----+----2----+----+----3----+----4----
NDIC1 = 'NDIC1.A'
NSEG = 'NSEG.A' ,NARB = 'NARB.A' ,NSEL = 'NSEL.A'
LISTP = OUI

PROC SELEC
=== SELECTION DES VARIABLES ===
NOMI ACT 1, 6--16
CONT ACT 2--5
FIN

PROC COLIG
=== TRANSPOSITION ===

PROC DISAR
=== PROCEDURE disar ===

PRCT = 33.33  var = 16  lord = 1  nxind = 20  >
nxniv = 13  lcout = lIST  lprob = prop  NRED = 5 >
NDIV = 5  nxnod = 200  ldesc = 1
ORDRE  10, 12
COUT   0.0   1.0   0.5>
       0.8   0.0   1.2>
       1.3   0.7   0.0

STOP
----+----1----+----2----+----+----3----+----4----
```

6. Fichiers nécessaires à l'exécution

- en lecture NDIC1 (fichier du dictionnaire utile apuré)
  NSEG (fichier des données transposées)

- en écriture NARB (fichier contenant l'arbre de décision binaire)
  NSEL (fichier des élagages)



Procédure DISAR :
La méthode

La procédure DISAR permet d'effectuer une **analyse discriminante non paramétrique**. Cette méthode constitue une alternative intéressante aux méthodes paramétriques de discrimination parmi lesquelles l'analyse linéaire discriminante et la régression logistique qui sont les plus utilisées. Ces dernières méthodes fournissent des règles de décision sous la forme d'expressions algébriques parfois difficiles à analyser et à interpréter.

L'analyse discriminante qui fait l'objet de la procédure DISAR est basée sur la construction d'un arbre de décision binaire obtenu à l'aide de divisions successives de sous-ensembles de l'échantillon en deux descendants. L'idée fondamentale est de sélectionner chaque division d'un sous-ensemble (ou segment) de telle sorte que les segments descendants soient plus "purs" que le segment parent. Autrement dit, il faut que le mélange des groupes soit moins important dans les segments descendants que dans le segment parent.

Un problème de discrimination se pose quand on est en présence d'un tableau de données contenant n sujets répartis en k classes C_1, C_2 ... C_k. Il s'agit alors d'une part, de sélectionner parmi les variables du tableau celles qui sont les plus discriminantes, et d'autre part, de construire une règle de décision permettant d'affecter un nouveau sujet à l'une de ces k classes.

DISAR est une approche de discrimination non paramétrique qui tient compte des interactions qui peuvent exister dans les données. Cette procédure est robuste vis-à-vis de données aberrantes, atypiques ou de sujets mal étiquetés. Elle est valable quelle que soit la nature des variables explicatives, sans codage préalable.

1. Construction d'un arbre de décision binaire

Considérons 300 sujets répartis en 3 classes C_1, C_2, C_3 de même taille et sur lesquels 10 mesures quantitatives V_1, V_2,... V_{10} ont été relevées. Au départ de la procédure de construction de l'arbre, on a un seul segment contenant les 300 sujets répartis en 3 classes de 100 sujets. DISAR passe alors en revue toutes les divisions possibles de la forme $V_1 < \alpha$ où α est une valeur quelconque contenue dans l'étendue de la première variable V_1. Chaque division scinde l'échantillon des 300 sujets en deux sous-ensembles ou segments descendants: le segment de gauche contient les sujets vérifiant $V_1 \leq \alpha$ et le segment de droite contient les autres sujets ($V_1 > \alpha$). De toutes ces divisions possibles de V_1, DISAR retient celle qui fournit les deux segments les moins mélangés en terme de séparation des classes. On peut, par exemple, aboutir à la division suivante qui n'est pas spécialement discriminante.

La même recherche de la meilleure division est effectuée avec toutes les autres variables V_2, ..., V_{10}. On obtient ainsi la meilleure division de chacune des 10 variables et l'algorithme retient finalement, parmi ces divisions, celle (par exemple $V_8 \leq 3{,}5$) qui fournit la meilleure séparation entre les 3 classes, ce qui peut se traduire par le dessin suivant :

Ensuite la même procédure de division est appliquée à chacun des deux segments descendants conduisant par exemple au diagramme suivant :

On pourrait arrêter là la procédure de division et produire l'arbre de décision binaire suivant à 4 segments terminaux, chacun d'eux étant en faveur de la classe qui y est la plus représentée (si les probabilités a priori sont estimées par les fréquences des classes dans l'échantillon, et si les coûts de mauvaise classification sont tous égaux à 1).

Mais cet arbre ne fournit pas une bonne règle de décision en terme d'erreur de classement. En effet, un sujet qui parcourt l'arbre et qui tombe dans le segment 1 est affecté à la classe 2 avec une erreur de classement de 14,86 %; celui qui tombe dans le segment 4 est affecté à la classe 2 avec une erreur de classement de 55,55 %. Le **T**aux d'**E**rreur **A**pparent de classement (TEA) associé à l'arbre est la moyenne des erreurs de classement dans les différents segments terminaux soit :

TEA = (74 x 14,86 % + 86 x 20,93 % + 95 x 26,31 % + 45 x 55,55 %) / 300 = 26,33 %

On a donc intérêt à continuer à diviser les segments. Mais quand s'arrêter ? Si on divisait jusqu'au bout on obtiendrait un arbre à 300 segments terminaux, chacun d'eux contenant un seul sujet, ce qui conduirait à un TEA égal à 0 qui est évidemment une estimation trop optimiste du **T**aux d'**E**rreur **T**héorique de classement (TET). En fait si on avait en réserve un échantillon de sujets tirés de la même population (échantillon test), on pourrait les faire courir le long de l'arbre à 300 segments terminaux et le taux d'erreur de classement de ces sujets serait peut être même plus grand que le TEA associé à l'arbre à 4 segments terminaux précédent. Le problème est donc de sélectionner, parmi la séquence de sous-arbres dont le nombre de segments terminaux va de 1 à 300, le "sous-arbre optimal", celui qui correspond à la plus petite valeur de l'estimation du TET.

C'est la solution proposée par L. Breiman et al.; elle constitue la base de la procédure DISAR et représente la référence actuelle en matière de technique d'arbre.

1.1 Prise en compte des variables qualitatives dans DISAR

Dans l'exemple illustratif, on a considéré des mesures, c'est-à-dire des variables quantitatives; mais les variables qualitatives peuvent également être prises en compte dans la procédure DISAR. Une variable binaire fournit une seule division. Une variable qualitative à k modalités ordonnées fournit (k-1) divisions, et, pour une variable qualitative à k catégories non ordonnées, l'algorithme examine toutes les divisions correspondant aux différents sous-ensembles de catégories et considère donc 2^{k-1} -1 divisions. Par exemple, si la variable présente 3 modalités non ordonnées m_1, m_2 et m_3, les divisions possibles sont au nombre de 3 :

à gauche, sujets de type m_1, à droite, sujets de type m_2 ou m_3

à gauche, sujets de type m_2, à droite, sujets de type m_1 ou m_3

à gauche, sujets de type m_3, à droite, sujets de type m_1 ou m_2

2. Sélection du "meilleur sous-arbre"

Comme il a été vu plus haut, un arbre petit (c'est-à-dire avec très peu de segments terminaux) entraîne un TEA qui, s'il estime correctement le TET, est trop important. Dans ce cas on peut être conduit à perdre de bonnes divisions et à ne pas utiliser toute l'information contenue dans l'échantillon. D'autre part, à un arbre très grand (avec de nombreuses divisions) est associé un TEA qui donne une estimation trop optimiste du TET. C'est donc entre ces deux extrêmes que doit être choisi le "meilleur" sous-arbre : celui dont le TEA est le plus petit possible tout en fournissant une estimation du TET la plus correcte possible. La recherche de ce "meilleur" sous-arbre dans DISAR se fait de la façon suivante :

- utilisation d'un **échantillon de base** pour construire un grand arbre A_{max} avec très peu de sujets dans chaque segment terminal. Le nombre de sujets, qui est un paramètre en option de la procédure (NXIND), est à fixer par l'utilisateur. Par exemple, on divise un segment tant que son effectif est supérieur à 5 (NXIND = 5). On peut cependant choisir NXIND = 1.

- utilisation d'un algorithme pour élaguer "judicieusement" les branches de ce grand arbre A_{max}. La procédure d'élagage produit une séquence optimale S de sous-arbres emboîtés de plus en plus petits telle qu'à chaque sous-arbre de cette séquence est associé le plus petit TEA comparé à celui de tout sous-arbre de même taille (c'est-à-dire ayant le même nombre de segments terminaux).

- le meilleur sous-arbre A^* est ensuite choisi parmi les sous-arbres de la séquence optimale S en utilisant soit la méthode de l'échantillon-test si la taille de l'échantillon est très grande soit la méthode de validation croisée si l'échantillon est de taille moyenne.

2.1 Méthode de l'échantillon-test

Dans cette méthode, il est nécessaire de disposer d'un **échantillon-test** ou de diviser l'échantillon total en deux parties :

(i) un échantillon de base qui est utilisé pour construire l'arbre A_{max} et déterminer la séquence optimale, et

(ii) un échantillon-test (1/3 de l'échantillon total en pratique; ce pourcentage est le paramètre PRCT de la procédure DISAR) qui sert à sélectionner le meilleur sous-arbre A^*. L'échantillon-test peut aussi être défini par la variable IVTES. Les sujets de l'échantillon-test parcourent alors chacun des sous-arbres de la séquence optimale et tombent dans un segment terminal, ce qui entraîne une estimation du TET pour chaque sous-arbre de S : c'est dans le cas le plus simple (probabilités a priori des classes égales aux fréquences et coûts de mauvaise classification égaux), la proportion des sujets de l'échantillon-test mal classés par le sous-arbre.

En pratique, cette estimation du TET est grande pour les grands sous-arbres de S; elle décroît quand la taille des sous-arbres diminue et croît ensuite quand les arbres deviennent trop petits. **L'arbre A^* sélectionné par DISAR est le plus petit sous-arbre de S associé à l'estimation la plus petite du TET**.

Remarque : Le programme DISAR fournit un intervalle de confiance associé à cette estimation TET. Dans le cas général les formules de calcul sont compliquées mais elles se simplifient si les probabilités a priori sont estimées par les fréquences des classes dans l'échantillon et si les coûts de mauvais classement sont pris tous égaux à 1. En effet : soient n_1 et n_2 les tailles respectives de l'échantillon de base et de l'échantillon test. L'estimation du TET est égale à la proportion p de sujets mal classés dans l'échantillon test et on sait résoudre le problème de l'estimation d'une proportion théorique p_t et de sa variance qui est donc égale ici à :

$$p_t\,(1-p_t)/n_2$$

Pour tenir compte des fluctuations au voisinage de l'estimation du TET minimum, on utilise dans DISAR la règle proposée dans Breinman et al. sous le nom de "1 s.e. rule" (règle d'un écart-type) dont le principe est le suivant :

Soit A^* l'arbre tel que $(A^*) =$

l'arbre sélectionné est celui qui a k_0 segments terminaux où (A_k) est la valeur la plus grande de (A_k) vérifiant :

$$(A_k) \leq (A^*) + \text{écart-type } [\ A^*)]$$

Remarque importante : L'appellation de taux d'erreur apparent ou théorique n'a de sens que si les probabilités a priori sont estimées par les fréquences des classes dans l'échantillon et si les coûts de mauvais classement sont tous égaux à 1. Dans le cas général il s'agit de **coût d'erreur apparent** ou **coût d'erreur théorique.**

2.2 Méthode de validation croisée

Si on ne dispose pas d'échantillon initial de taille importante, il est fortement conseillé d'utiliser la méthode de validation croisée pour estimer les TET associés aux sous-arbres de la séquence S.

L'échantillon total L est divisé aléatoirement en V sous-échantillons de taille quasiment égales et mutuellement exclusifs L_v (v = 1, ... , V). On prend en général V égal à 10.

L'arbre A_{max} est construit à l'aide de l'échantillon total L, et V autres arbres auxiliaires sont construits à l'aide des V échantillons notés L^v, complémentaires de L_v dans L. Les échantillons L_v qui n'ont pas contribué à la construction des arbres peuvent être utilisés comme échantillons-test et fournir des estimations sans biais des TET associés aux sous-arbres des arbres . La procédure d'élagage permet de construire la séquence S = de sous-arbres emboîtés associée à L ainsi que V séquences S_v = associées à L^v. Pour sélectionner le sous-arbre A^* le plus fiable parmi ceux de S, on cherche quel est pour chaque v (v = 1, ... , V) le sous-arbre de la séquence S_v le plus proche de chaque sous-arbre de S.

Breiman et al proposent d'utiliser les valeurs du paramètre de complexité α. Ainsi, pour chaque sous-arbre A_h de S on définit la quantité , moyenne géométrique de α_h et α_{h+1}, car c'est pour α compris entre ces deux valeurs que l'arbre A_h est l'arbre $A(\alpha)$, plus petit sous-arbre ayant une mesure de coût-complexité minimale.

On choisit alors dans S_v le sous-arbre A_{iv} dont la valeur du paramètre de complexité α_{iv} est inférieure ou égale à tout en étant la plus grande possible. A chaque L^v est associée une règle d'affectation RA(v) qui permet de classer les individus de L_v parcourant les sous-arbres A_{iv}. On obtient ainsi, pour chaque sous-arbre de la séquence A_h de la séquence S, une estimation du TET. Et le sous-arbre A* à sélectionner est celui qui correspond à l'estimation du TET la plus petite.
Le lecteur pourra se reporter aux ouvrages L. Breiman et al et G. Celeux et J.P. Nakache pour une justification théorique de la méthode de validation croisée et à l'article A. Gueguen et J.P. Nakache pour une application numérique de cette méthode.

Remarque : il est plus difficile d'obtenir un écart-type de l'estimation du TET quand la validation croisée est utilisée : en effet il n'y a pas d'indépendance entre les échantillons L_v et L^v. Cepedant, si on accepte en pratique de ne pas tenir compte de cette non-indépendance comme le font Breiman et al, on obtient une formule semblable à celle obtenue dans la méthode de l'échantillon-test.

3. Autres caractéristiques de DISAR

3.1 Probabilités a priori, coûts de mauvaise classification

La règle de décision la plus générale est celle qui tient compte des probabilités a priori π_j (j = 1,2, ... k) des k classes à discriminer et des coûts de mauvaise classification C(j/s) :

$$C(j/s) \qquad (j \neq s = 1,2,...,k)$$

C(j/s) le coût entraîné par l'affectation d'un individu au groupe Cj, alors qu'en réalité il appartient au groupe Cs ($s \neq j$). Les différents coûts C(s/s) sont nuls et, en général,

$$C(j/s) \neq C(s/j)$$

Une option de la procédure DISAR permet de spécifier ces probabilités (LPROB) et ces coûts (LCOUT). On suppose souvent en pratique ces coûts égaux et on estime les probabilités a priori des groupes par les fréquences des différents groupes dans l'échantillon.

La règle générale d'affectation d'un segment terminal à une classe est basée sur le coût moyen d'erreur de classement (ou risque d'erreur). Si on considère le segment terminal T, il contient $n_1(T)$ sujets de classe 1, $n_2(T)$ sujets de la classe 2, ..., $n_j(T)$ sujets de la classe j, ... $n_k(T)$ sujets de la classe k. n_j étant l'effectif total de la classe j on a :

$$P(j/T) = P(j,T) / P(T)$$
$$P(j,T) = \pi_j \, nj(T) / nj$$

$$P(T) = \sum_{j=1}^{k} P(j,T) \qquad \text{probabilité d'arriver dans T.}$$

Le coût moyen d'erreur de classement, noté R(s/T), entraîné par l'affectation du segment T à la classe C_s est égal à :

$$R(s/T) = \sum_{j=1}^{k} C(s/j) \, P(j/T)$$

Ainsi le segment terminal T est affecté à la classe C_r si :

$$R(r/T) = \min \{R(s/T); s = 1,2,...k\}$$

Remarque : Si $\pi_j = n_j / n$, proportion d'observations de Cj dans l'échantillon alors :

$P(T) = \sum_{j=1}^{k} n_j (T) / n$ = proportion de sujets allant dans le segment terminal T.

Remarque : les probabilités a priori et les coûts de mauvais classement interviennent dans la construction de l'arbre, dans l'élagage, dans le calcul des coûts d'erreur et de l'écart-type de l'estimation du coût d'erreur théorique, et enfin dans la détermination des variables concurrentes et suppléantes.

3.2 Divisions équi-réductrices (ou concurrentes), divisions équi-divisantes (ou suppléantes)

Dans la description des divisions de l'arbre A^*, il est question de divisions équi-réductrices et de divisions équi-divisantes. Leurs nombres respectifs (NRED, NDIV) sont des paramètres de la procédure DISAR dont les valeurs sont à fixer par l'utilisateur.

- **La première division équi-réductrice (ou concurrente)** d'un segment t est celle qui correspond à une réduction de l'impureté la plus proche de celle de la meilleure division d^*. C'est en fait la deuxième meilleure division du segment. On définit ainsi les 2ème, 3ème, ..., divisions équi-réductrices .
  La meilleure division d* est obtenue comme suit : on cherche **pour chaque variable** V_j la meilleure division d^*_j qui assure la réduction de l'impureté maximum. Si l'on note la réduction de l'impureté pour la division de la variable , est alors la division telle que :

  .

  La division choisie est effectuée à l'aide de la variable dont la division assure: où p est le nombre de variables explicatives.

  La 1ère division équi-réductrice est la division d^*_i effectué sur la variable V_j telle que :

- **Les divisions équi-divisantes (ou suppléantes)** permettent de classer un nouveau sujet présentant des données manquantes. L'idée est la suivante : si la meilleure division d* du segment t est obtenue à partir de la variable V_j, on définit pour chaque variable V_i avec i = 1,2,..., p où p est le nombre de variables explicatives et $i \neq j$, la division d_i, la plus semblable à d^* : d_i est une division équi-divisante de d*. La meilleure division équi-divisante est, parmi les (p-1) divisions équi-divisantes d_i ($i \neq j$), la division équi-divisante la plus semblable à d^*. De la même manière, on peut définir la seconde (meilleure) division équi-divisante (ou suppléante), la troisième, etc.

  Remarque : Il est possible que le programme ne fournisse pas de division suppléante bien que la demande d'un certain nombre de ces divisions soit faite dans le fichier de commande. Cette situation peut s'expliquer par l'exemple suivant où, pour une raison de simplicité, les probabilités a priori sont estimées par les fréquences des groupes dans l'échantillon.

On considère la meilleure division et la division **triviale** suivantes :

où les nombres indiqués dans les segments sont les effectifs de ces segments. Le tableau de contingence issu du croisement de avec est le suivant :

20	10	30
0	0	0
20	10	30

On remarque alors que dans 2/3 des cas cette division prédit correctement . Et donc toute division qui fournirait une prédiction de inférieure à celle fournie par la division ne peut être considérée comme suppléante. C'est le cas, par exemple, de la division suivante :

11	4	15
9	6	15
20	10	30

Dans DISAR, la mesure d'association entre d_m et d^* serait négative. Les divisions suppléantes correspondent à des valeurs de cette mesure variant entre 0 (pas d'association) et 1 (association parfaite).

Procédure DISEL :
Elagage de l'arbre pour la discrimination

1. Présentation

1.1 Objet

La procédure DISEL est exécutable après DISAR, c'est-à-dire après construction d'un arbre de segmentation binaire utilisé pour discriminer k groupes.

Cette procédure effectue l'édition détaillée d'un sous-arbre de la séquence d'élagage, à partir de la lecture des fichiers NARB et NSEL.

1.2 Editions

Edition du dendogramme de l'arbre, suivi d'une description succincte des divisions. Edition détaillée des noeuds de l'arbre et écriture de la règle d'affectation aux différents groupes.

1.3 Paramètres

Il y a deux paramètres dans cette procédure : le type d'édition des individus dans les segments terminaux (LEDIT) et le nombre de segments terminaux choisis (NOTER). Le choix de ce dernier paramètre détermine l'arbre à éditer.

Attention : ce paramètre doit prendre une des valeurs listées à la fin de l'exécution de DISAR sous le titre "Segments terminaux" (ce sont les seuls arbres qu'il soit possible de construire).

2. Instructions de commande

Les valeurs par défaut sont indiquées entre parenthèses, à la suite du nom du paramètre. Si on désire que les paramètres prennent leurs valeurs par défaut, on codera le mot-clef NOPAR à la place de la liste des paramètres repérés par (**3**).

(1) PROC DISEL Edition d'un arbre de segmentation

(2) *titre de la procédure*

(3) NOTER (0) nombre de segments terminaux de l'arbre à éditer.

LEDIT (0 ou NON) édition de la correspondance segments terminaux - individus.

3. Présentation détaillée des paramètres

NOTER **nombre de segments terminaux de l'arbre à éditer**

- *valeurs possibles* : une des valeurs du tableau d'élagage (proc DISAR)
- *valeur par défaut* : 0

La procédure DISAR qui précède a édité le tableau d'élagage. NOTER doit être l'une des valeurs de la colonne "segments terminaux". L'arbre édité aura NOTER segments terminaux.

Si NOTER = 0 et si PRCT ou IVTES est différent de 0 dans DISAR, NOTER prendra la valeur repérée par une étoile dans le tableau d'élagage, c'est-à-dire celle qui correspond à l'arbre optimum.

Remarque : Il arrive que l'arbre optimum ne comporte qu'un seul segment (l'échantillon n'est pas divisible). Ceci correspond en général à une discrimination difficile à réaliser avec les variables choisies. L'utilisateur peut

Cependant dans ce cas choisir un arbre sous-optimal dans la liste fournie par DISAR.

LEDIT **édition de la correspondance segments terminaux - individus**

- *valeurs possibles* :
  - 0 ou NON pas d'édition
  - 1 ou COMPO composition de chaque groupe du nœud terminal
  - 2 ou AFFEC groupe d'appartenance de chaque individu par nœud terminal
  - 3 ou TOT options COMPO + AFFEC
- *valeur par défaut* : 0 ou NON

On peut désirer connaître de façon précise la répartition des individus dans les groupes d'origine des nœuds terminaux. Deux types d'éditions sont proposées ici.

- Si LEDIT = COMPO, on obtiendra pour chaque nœud terminal et pour chaque groupe d'origine, la liste des individus qui en font partie.

- Si LEDIT = AFFEC, on donne la liste des individus appartenant au segment terminal, avec pour chacun d'eux le numéro du groupe d'origine auquel il appartient.

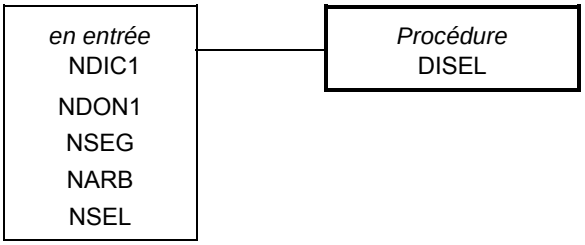
Attention : dans cette sortie, les individus sont repérés par leur identificateur court (en 4 caractères).

4. Exemple

```
----+----1----+----2----+----+----3----+----4----
NDIC1 = 'NDIC1.A' NDON1 = 'NDON1.A'
NSEG = 'NSEG.A' ,NARB = 'NARB.A' ,NSEL = 'NSEL.A'
LISTP = OUI
PROC DISEL
exemple de procédure disel
NOTER = 0 LEDIT = 1
STOP
----+----1----+----2----+----+----3----+----4----
```

5. Fichiers nécessaires à l'exécution

• en lecture NDIC1 (fichier dictionnaire utile apuré)
NDON1 (fichier des données utiles apuré)
NSEG (fichier des données transposées)
NARB (fichier contenant l'arbre de décision binaire)
NSEL (fichier des élagages)



Procédure REGAR : *Régression non paramétrique par arbre de segmentation*

1. Présentation

1.1 Objet

Cette procédure effectue la construction d'un arbre de décision binaire complet pour la régression non-paramétrique d'une variable à expliquer quantitative (variable réponse) sur un ensemble de variables explicatives qui peuvent être de nature quelconque : continues, ordinales ou nominales.

La méthode fournit, à partir de l'arbre binaire complet, la séquence des sous-arbres obtenue en utilisant une procédure d'*élagage* basée sur la suppression successive des branches les moins informatives en termes d'explication de la variable réponse Y. Dans cette séquence d'*élagage*, elle sélectionne un sous-arbre "optimal" à l'aide de *l'échantillon de base* (en se basant sur l'estimation de la variance résiduelle des différents segments terminaux).

Les variables explicatives ne doivent présenter aucune donnée manquante. Les données déclarées "manquantes" dans la variable continue à expliquer repèrent en réalité les individus dits "anonymes".

La procédure REGAR doit être précédée de la procédure de gestion COLIG utilisée pour transposer le tableau de travail. La procédure utilise en entrée les fichiers de type NDIC et NSEG.

1.2 Editions

La procédure permet de tracer l'arbre. Cette édition n'est pas faite si le nombre de niveaux est supérieur à 13. Elle donne, quel que soit le nombre de niveaux, une description succincte : description des divisions et taille des segments.

Le programme fournit en option la séquence d'élagage, et pour tout arbre de cette séquence, l'ensemble de ses segments terminaux.
Enfin le programme édite les informations qui permettent de déterminer le sous-arbre optimal : à chaque arbre de la séquence est associée l'estimation de la variance résiduelle rapportée à la variance totale de l'échantillon de base, et cette même estimation (avec son écart-type) pour l'échantillon-test.

1.3 Paramètres

Les paramètres de la procédure se divisent en trois catégories:

- **Les paramètres de définition de l'analyse**: la variable à expliquer (VAR), la présence de variables ordinales (LORD); le pourcentage d'individus dans l'échantillon-test (PRCT), ou l'identification de la variable servant à définir un échantillon-test (IVTES).

- **Les paramètres de fonctionnement** permettent de fixer les caractéristiques de l'arbre binaire : effectif minimum requis pour diviser un segment (NXIND); nombre maximum de niveaux de l'arbre (NXNIV); nombre maximum de segments de l'arbre (NXNOD).

- **Les paramètres d'édition**: nombre de divisions "equi-réductrices" (ou concurrentes) (NRED), nombre de divisions "equi-divisantes" (ou suppléantes) (NDIV) ; la description complète de l'arbre (LDESC); le dessin de l'arbre complet (LDESS); la description de la séquence d'élagage (LSEQU).

1.4 Commandes par listes

Si on a déclaré des variables ordinales (LORD = OUI), la liste de ces variables devra suivre **immédiatement**. Elle sera écrite avec le format approprié, et ne contiendra ni ligne vide ni commentaire (il s'agit de **données** et non d'instructions).

2. Instructions de commande

Les valeurs par défaut sont indiquées entre parenthèses, à la suite du nom du paramètre.

(1) PROC REGAR — Régression par arbre de segmentation

(2) — *titre donné à l'analyse*

(3) VAR — numéro de la variable continue à expliquer (pas de valeur par défaut).

PRCT (33.0) — pourcentage d'individus dans l'échantillon-test.
obligatoire, PRCT doit être différent de zéro sauf si IVTES ≠ 0.

IVTES (0) — numéro d'une variable servant à définir un échantillon-test

LORD (0 ou NON) — présence d'une liste de variables ordinales.
0 ou NON : pas de variables ordinales.
1 ou OUI : il y a une liste de variables ordinales.

NXIND (5) — nombre minimum d'individus dans un segment divisible.

NXNIV (13) — nombre maximum de niveaux de l'arbre.

NXNOD (200) — nombre maximum de segments.

NRED (0) — nombre de divisions equi-réductrices (ou concurrentes) à éditer.

NDIV (0) nombre de divisions equi-divisantes (ou suppléantes) à éditer.
LDESC (0 ou NON) description de l'arbre complet.
1 ou OUI : édition de l'arbre.
0 ou NON : pas d'édition.
LDESS (0 ou NON) dessin de l'arbre complet
LSEQU (0 ou NON) description de la séquence d'élagage.

(4) Liste des variables ordinales (si LORD = OUI)

3. Présentation détaillée des paramètres

VAR **numéro de la variable continue à expliquer**

- *valeurs possibles* : entières de 1 à NQEXA
- ***pas de valeur par défaut***

La variable à expliquer est une variable continue. Une valeur manquante pour cette variable (repérée par le code donnée manquante déclaré dans ARDON) correspond à un individu anonyme dont il faudra prédire la valeur.

PRCT **pourcentage d'individus dans l'échantillon - test**

- *valeurs possibles* : réelles non nulles
- *valeur par défaut* : 33.0

PRCT représente le pourcentage d'individus tirés au hasard dans l'échantillon total. Le reste des individus, appelé échantillon de base, sert à construire l'arbre complet et la séquence des sous-arbres.
Dans la méthode de segmentation, l'échantillon-test sert à sélectionner un sous-arbre "optimal".

Remarque : L'échantillon-test est constitué en coupant l'étendue de la variable à expliquer en k intervalles et en tirant au hasard PRCT % d'individus dans chacun de ces intervalles.

IVTES **numéro de la variable servant à définir un échantillon-test**

- *valeurs possibles* : réelles
- *valeur par défaut* : 0

Cette variable servant à définir l'échantillon-test peut prendre deux valeurs:
- **1** pour repérer les individus appartenant à **l'échantillon-test**

• **2** pour repérer les individus appartenant à **l'échantillon de base**

Au sens de SPAD.N, c'est donc une variable nominale à 2 modalités.
Si IVTES = 0, on n'utilise pas de variable pour définir l'échantillon-test. Dans ce cas, l'échantillon-test est défini par PRCT :

* Si PRCT = 0, on ne fera pas d'élagage de l'arbre.
* Si PRCT > 0, l'échantillon-test sera défini par tirage aléatoire.

Si IVTES > 0, indiquez le numéro de la variable permettant de définir l'échantillon-test.

La définition d'un échantillon-test est obligatoire dans cette méthode. Cet échantillon peut être défini par le paramètre IVTES, ou par le paramètre PRCT (cf. sa définition). IVTES est prioritaire.

LORD **présence d'une liste de variables nominales à considérer comme ordinales**

• *valeurs possibles* : 0 ou NON pas de variables ordinales
1 ou OUI présence d'une liste de variables ordinales

• *valeur par défaut* : 0 ou NON

Dans la méthode de segmentation, les variables ordinales jouent un rôle particulier. Si LORD = 1 ou OUI, la liste des variables ordinales est introduite par le mot clé ORDRE après la liste des paramètres généraux. Les variables doivent avoir été sélectionnées comme variables nominales dans la procédure SELEC qui a précédé.

Conseils

• Si le problème met en jeu plusieurs variables nominales ayant un grand nombre de modalités, le nombre de divisions dans la procédure est très important.
Par exemple, pour une variable nominale à 10 modalités, le nombre de divisions sera égal à $2^{10-1} - 1 = 511$; ce nombre est égal à 16383 si la variable nominale possède 15 modalités. Le programme utiliserait donc dans ce cas un volume important de mémoire. En cas de nécessité, on procédera à des regroupements de modalités.

• Lorsque l'on a des variables ordinales avec des réponses manquantes, et que l'on a codé REC dans SELEC, les réponses manquantes ont la valeur n + 1, et seront donc du côté des valeurs les plus fortes.

NXIND **nombre minimum d'individus dans un segment divisible**

• *valeurs possibles* : entières positives

• *valeur par défaut* :5

Un segment n'est plus divisé si l'effectif qui se trouve est inférieur à NXIND.

Le choix NXIND = 1 conduit à des arbres de grande taille, pouvant avoir autant de segments terminaux que d'individus.
La valeur par défaut convient en général. NXIND est un des paramètres qui arrête la division des segments.

Remarque: le segment concerné pouvant être divisé, il sera normal de trouver des segments terminaux où l'effectif est inférieur à NXIND

NXNIV **nombre maximum de niveaux de l'arbre**

- *valeurs possibles* : entières de 1 à 32
- *valeur par défaut* :13

Le nombre de niveaux d'un arbre se calcule de la façon suivante : le premier segment est au niveau 0 ; ses deux descendants immédiats sont au niveau 1, etc... La valeur du paramètre NXNIV indique le nombre maximum de niveaux que peut atteindre "l'arbre maximum". NXNIV est un des paramètres qui arrête la division des segments.

Remarque • Si un arbre a plus de 13 niveaux, il est construit mais il n'est pas dessiné.

NXNOD **nombre maximum de segments**

- *valeurs possibles* : entières positives
- *valeur par défaut* : 200

La construction de l'arbre maximum s'arrête dès que le nombre de segments est égal à NXNOD. NXNOD est un paramètre de contrôle qui arrête la division des segments.

Attention : Si le nombre total de segments édité dans la description de l'arbre complet avant élagage est voisin de 200 (valeur par défaut), l'arbre a beaucoup de chance de ne pas être complètement construit. Dans ce cas il faut relancer la procédure en augmentant la valeur de NXNOD.

NRED **nombre de divisions equi-réductrices (ou concurrentes) à éditer**

- *valeurs possibles* : entières positives ou nulles
- *valeur par défaut* :0

La meilleure division est celle qui assure la plus grande réduction de la variance résiduelle en passant d'un segment parent à ses segments descendants.
Par abus de langage, on appelle "equi-réductrices" ou "concurrentes" les divisions qui assurent les plus fortes réductions de la variance résiduelle des segments descendants. Par exemple si NRED = 2 on obtient les deux meilleures divisions après la division fournie (qui est en réalité la meilleure).
Ce paramètre est utilisé au moment de l'édition de la description des divisions. Au maximum NRED divisions equi-réductrices sont éditées. En pratique, deux à cinq divisions equi-réductrices sont en général suffisantes.

NDIV **nombre de divisions equi-divisantes (ou suppléantes) à éditer**

- *valeurs possibles* : entières positives ou nulles
- *valeur par défaut* : 0

Par extension on appelle "equi-divisantes" ou "suppléantes" les divisions qui fournissent les répartitions les plus proches de la meilleure division.
Si NDIV = 2 on obtient les deux divisions qui assurent les répartitions les plus proches de la répartition considérée.
Ce paramètre est utilisé au moment de l'édition de la description des divisions. Au maximum NDIV divisions equi-divisantes sont éditées. En pratique, deux à cinq divisions equi-divisantes sont en général suffisantes.

Le code « c » à côté des variables équi-divisantes indique qu'il faut inverser les règles d'affectation.

LDESC **description de l'arbre complet**

- *valeurs possibles* : 0 ou NON pas de description
  1 ou OUI description de l'arbre
- *valeur par défaut* : 0 ou NON

Si LDESC = OUI, on obtient l'édition de l'arbre complet. Cette édition peut être volumineuse et l'arbre difficile à lire. L'étape REGEL (choix de l'arbre élagué) permet l'édition du sous-arbre choisi.

LDESS **dessin de l'arbre complet**

- *valeurs possibles* : 0 ou NON pas de dessin de l'arbre
  1 ou OUI dessin de l'arbre
- *valeur par défaut* : 0 ou NON

NB : L'arbre existe toujours mais n'est pas édité s'il possède plus de 13 niveaux.

LSEQU **description de la séquence d'élagage**

- *valeurs possibles* : 0 ou NON pas de description
  1 ou OUI description de la séquence d'élagage.
- *valeur par défaut* : 0 ou NON

Remarques générales

• Si le problème met en jeu plusieurs variables nominales ayant un grand nombre de modalités, le nombre de divisions dans la procédure est très important.
Par exemple, pour une variable nominale à 10 modalités, le nombre de divisions sera égal à $2^{10-1} - 1 = 511$; ce nombre est égal à 16383 si la variable nominale possède 15 modalités. Le programme utiliserait donc dans ce cas un volume important de mémoire. En cas de nécessité, on procédera à des regroupements de modalités.

• Lorsque l'on a des variables ordinales avec des réponses manquantes, et que l'on a codé REC dans SELEC, les réponses manquantes ont la valeur n + 1, et seront donc du côté des valeurs les plus fortes. Il est recommandé de se débarrasser des valeurs manquantes.

4. Listes de sélection des variables ordinales

Après la liste de paramètres, il faut insérer la liste des variables ordinales (si LORD = OUI).

La syntaxe à employer suit le schéma suivant:

mot-clé *liste de numéros*

• **mot-clé** imposé est : ORDRE.

• *Liste de numéros* est une liste de nombres donnant les numéros d'origine des variables ordinales choisies. Le séparateur entre les nombres peut être le blanc ou la virgule.

Une telle liste est une instruction SPAD : on peut donc utiliser le caractère de continuation ">" ainsi que le symbole de suite "--" .

Tous les numéros présents dans la liste des variables ordinales doivent être entiers et doivent correspondre à des variables nominales de NDON (donc de NSEG). Ces nombres seront donc inférieurs ou égaux au nombre total de variables (NQEXA).

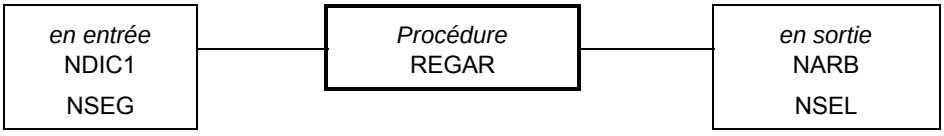
5. Exemple

```
----+----1----+----2----+----+----3----+----4----
NDIC1 = 'NDIC1.A'
NSEG = 'NSEG.A' ,NARB = 'NARB.A' ,NSEL = 'NSEL.A'
LISTP = OUI

PROC SELEC
=== SELECTION DES VARIABLES ===
NOMI ACT 20---23,31
CONT ACT 17,18
FIN

PROC COLIG
=== TRANSPOSITION ===

PROC REGAR
=== PROCEDURE REGAR ===

var = 17, PRCT = 33.33, IVTES = 0 >
nxind = 20, nxniv = 13, nxnod = 200 >
NRED = 5, NDIV = 5, ldesc = OUI, lord = OUI
ORDRE  20--23, 31

STOP
----+----1----+----2----+----+----3----+----4----
```

6. Fichiers nécessaires à l'exécution

- en lecture NDIC1 (fichier du dictionnaire utile apuré)
  NSEG (fichier des données transposées)

- en écriture NARB (fichier contenant l'arbre de décision binaire)
  NSEL (fichier des élagages)



PROCEDURE REGAR :
La méthode

La procédure REGAR permet d'effectuer une régression multiple par arbre de prédiction binaire. Elle constitue une alternative à la méthode paramétrique usuelle (régression linéaire multiple) qui fournit une règle de prédiction sous forme de combinaison linéaire des variables explicatives.

La méthode de régression qui fait l'objet de la procédure REGAR est basée sur la construction d'un arbre binaire. Cet arbre est obtenu à l'aide de divisions successives de sous-ensembles de l'échantillon en deux descendants.

L'idée fondamentale est de sélectionner chaque division d'un sous-ensemble (ou segment) de telle sorte que la variance de la variable à expliquer Y dans les segments descendants soit plus faible que la variance de Y dans le segment parent. Plus précisément, pour toute division d'un segment, on calcule la moyenne pondérée de la variance de Y dans ses segments descendants (variance *intra*). La meilleure division est, parmi toutes les divisions possibles à l'aide des variables explicatives, celle qui minimise cette variance-*intra.* On retient ainsi la division qui conduit à deux segments descendants aussi homogènes que possible en Y.

Un problème de régression multiple se pose quand on est en présence d'un tableau de données contenant une variable privilégiée (variable réponse), Y continue **à expliquer** par les autres variables du tableau X_1, X_2 ... X_p dites variables **explicatives**. Il s'agit alors d'une part, de sélectionner parmi ces variables celles qui expliquent significativement le phénomène Y, et d'autre part, de construire une règle de prédiction de la valeur de Y pour un nouvel individu.

REGAR est une approche de **régression non paramétrique** qui tient naturellement compte des interactions qui peuvent exister dans les données. Cette procédure est, d'autre part, robuste vis-à-vis de données aberrantes ou atypiques. Elle est valable quelle que soit la nature des variables explicatives analysées sans aucun codage préalable (continues, nominales, ordinales).

1. Construction d'un arbre de décision binaire

Considérons un ensemble d'individus sur lesquels on relève les informations concernant une variable continue Y à expliquer et 10 variables explicatives continues V_1, V_2,.. V_{10}.
Au départ de la procédure de construction de l'arbre, on a un seul segment contenant l'ensemble des individus. On suppose que les valeurs de Y ont pour moyenne m = 10 et pour variance s^2 = 60.
Le programme passe en revue toutes les divisions possibles de la forme $V_1 < \alpha$ où α est une valeur quelconque contenue dans l'étendue de la première variable V_1.

Chaque division scinde l'échantillon initial en deux sous-ensembles ou segments descendants : le segment de gauche contient les individus vérifiant $V_1 \leq \alpha$ et le segment de droite contient les autres individus ($V_1 > \alpha$). De toutes ces divisions possibles de V_1, on retient celle qui fournit les deux segments les plus homogènes (les moins dispersés) possible en Y. On peut, par exemple, aboutir à la division suivante :

(Cette meilleure division obtenue avec la variable V_1 ne semble pas efficace en terme de réduction de la variance à l'intérieur des segments descendants.)

La même recherche de la meilleure division est effectuée successivement avec toutes les autres variables V_2, ..., V_{10}. On obtient ainsi la meilleure division pour chacune des 10 variables et on retient finalement la division pour laquelle

la moyenne pondérée des variances de Y (dans les deux segments descendants) est minimum. Soit : par exemple $V_5 \leq 7{,}2$.

La même procédure est ensuite appliquée à chacun des deux segments descendants obtenus.

Si, parmi les variables explicatives, certaines sont qualitatives, elles sont prises en compte dans les calculs de la manière suivante : une variable nominale à deux modalités fournit une seule division. Une variable nominale à k modalités ordonnées fournit (k-1) divisions, et, pour une variable nominale à k modalités non ordonnées, le programme examine toutes les divisions correspondant aux différents sous-

ensembles de modalités et considère donc 2^{k-1} -1 divisions. Par exemple si la variable présente 3 modalités non ordonnées m_1, m_2 et m_3, les divisions possibles sont au nombre de 3 :

à gauche, individus de type m_1; à droite, individus de type m_2 ou m_3
à gauche, individus de type m_2; à droite, individus de type m_1 ou m_3
à gauche, individus de type m_3; à droite, individus de type m_1 ou m_2

On peut aboutir à l'arbre à deux niveaux suivant :

On pourrait arrêter là la procédure de division et produire l'arbre de prédiction à 4 segments terminaux suivant :

Considérons un nouvel individu, dont on cherche à prédire la valeur de Y. Il tombera dans un de ces 4 segments terminaux après avoir parcouru un chemin de l'arbre suivant les valeurs qu'il présente pour V_5, V_{11} et V_{16}. On obtient ainsi une valeur

prédite de Y (la moyenne dans le segment) et un écart-type (l'écart-type dans le segment).

Si certaines variances des segments sont encore importantes, on peut préférer continuer de diviser dans le but de réduire davantage les variances des segments terminaux.
Plus on divise, plus les variances décroissent, pour être finalement nulles quand chaque segment terminal contient un seul individu. Au grand arbre complet A_{max} ainsi obtenu est affectée une *erreur apparente de prévision* (EAP) nulle. On est alors confronté au problème de la détermination d'un sous-arbre "optimal" le moins grand possible et qui fournit l'estimation la plus petite de *l'erreur théorique de prévision* (ETP) :

1.1 Erreur Apparente de Prévision (EAP) associée à un arbre A

A chaque segment terminal t de l'arbre A est associée l'erreur suivante :

$$R(t) = p(t) * s^2(t)$$

avec et

où : N = nombre total d'individus (taille du segment origine)
N (t) = nombre d'individus du segment t
moyenne des valeurs Y des individus du segment t

L'erreur apparente de prévision associée à l'arbre A a pour expression :

EAP(A) =

EAP(A) est la moyenne pondérée des variances de Y dans chacun des segments terminaux de l'arbre A. Le rapport EAP(A) / s^2, où s^2 est la variance de Y à la racine, est l'équivalent de l'expression (1 - R^2) de la régression linéaire multiple (qui correspond au pourcentage de la variance totale inexpliquée par les covariables).

Remarque : dans la régression linéaire multiple, on suppose que la variance de la réponse Y conditionnellement aux covariables est constante, ce qui n'est pas le cas dans la régression par arbre.

2. Sélection du "meilleur" sous-arbre

Un arbre petit (c'est-à-dire avec très peu de segments terminaux) entraîne une Erreur Apparente de Prédiction (EAP) qui, si elle estime correctement l'Erreur Théorique de Prédiction (ETP), est trop importante. Dans ce cas on peut être conduit à perdre de bonnes divisions et à ne pas utiliser toute l'information contenue dans l'échantillon. D'autre part à un arbre très grand (avec de nombreuses divisions) est associé une EAP qui donne une estimation trop optimiste de l'ETP. C'est donc entre ces deux extrêmes que doit être choisi le "meilleur" sous-arbre.

La recherche de ce "meilleur" sous-arbre se fait de la façon suivante :

- utilisation d'un **échantillon de base** pour construire un grand arbre A_{max} tel que chacun des segments terminaux contienne peu d'individus. Ce nombre d'individus, qui est un paramètre en option de la procédure (NXIND), est à fixer par l'utilisateur. Par exemple si NXIND =5, on divisera tout segment tant que son effectif est supérieur ou égal à 5.

- utilisation d'un algorithme pour "élaguer" judicieusement les branches de ce grand arbre A_{max} en supprimant successivement les branches les moins informatives en terme d'explication de la variance de Y. La procédure d'élagage produit une séquence S^* de sous-arbres emboîtés de plus en plus petits telle qu'à chaque sous-arbre de cette séquence est associée la plus petite EAP comparée à celle de tout sous-arbre de même taille (c'est à dire ayant le même nombre de segments terminaux).

- le meilleur sous-arbre A^* est ensuite choisi parmi les sous-arbres de la séquence optimale S^* à l'aide d'un échantillon-test. Par conséquent **il est absolument nécessaire de disposer d'un échantillon-test** ou de diviser l'échantillon total en deux parties :
  (i) un échantillon de base qui est utilisé pour construire l'arbre A_{max} et déterminer la séquence optimale S^*, et
  (ii) un échantillon-test (1/3 de l'échantillon total en pratique ; ce pourcentage est un paramètre PRCT de la procédure REGAR) qui sert à sélectionner le meilleur sous-arbre A^*.
  L'échantillon-test peut aussi être défini par la variable IVTES.

  Les individus de l'échantillon-test parcourent alors chacun des sous-arbres de la séquence optimale et tombent dans un segment terminal, ce qui entraîne une estimation de l'ETP pour chaque sous-arbre de S^*.

En pratique, l'estimation de l'ETP décroît rapidement à mesure que le nombre de segments terminaux des sous-arbres augmente, puis elle passe par un plateau et croît ensuite lentement. **Le meilleur arbre A^* sélectionné comme optimal est le plus petit sous-arbre de S^* associé à l'estimation la plus petite de l'ETP**.

Calcul de l'estimation de l'erreur théorique pour un sous-arbre A de la séquence S^*

Cette estimation est obtenue à l'aide des valeurs de la variable à expliquer des individus de l'échantillon-test dont la taille est N^{et}.
A chaque segment t terminal de l'arbre est associée l'erreur suivante :

avec

où est le nombre d'individus de l'échantillon-test qui tombent dans le segment t

et

où est la moyenne prédite de y pour le segment t.

On en déduit

Remarque : Pour tenir compte des fluctuations au voisinage de l'estimation de l'ETP minimum, on utilise la règle connue dans l'ouvrage CART sous le nom de "1 s.e. rule", (règle d'un écart-type), dont le principe est le suivant :

Soit A^* l'arbre tel que

l'arbre sélectionné est celui qui a k_0 segments terminaux où est la valeur la plus grande de vérifiant :

3. Autres caractéristiques du programme

3.1 Divisions equi-réductrices (ou concurrentes), divisions equi-divisantes (ou suppléantes)

Dans la description des divisions de l'arbre A^*, il est question de divisions équi-réductrices (ou concurrentes) et de divisions équi-divisantes (ou suppléantes). Leurs nombres respectifs (NRED, NDIV) sont des paramètres de la procédure REGAR dont les valeurs sont à fixer par l'utilisateur.

- **La première division équi-réductrice (ou concurrente)** d'un segment t est celle qui correspond à une réduction de la variance résiduelle des segments descendants la plus proche de celle de la meilleure division d^*. C'est en fait la deuxième meilleure division du segment t. On définit ainsi les 2ème, 3ème, ..., divisions équi-réductrices.
  La meilleure division d* est obtenue comme suit : on cherche **pour chaque variable** V_j la meilleure division d^*_j qui assure la réduction de la variance

résiduelle maximum. Si l'on note la réduction de la variance résiduelle pour la division de la variable , est alors la division telle que : .

La division choisie est effectuée à l'aide de la variable dont la division assure :

où p est le nombre de variables explicatives.

La 1ère division ou variable équi-réductrice est la division d^*_i effectuée sur la variable V_j telle que :

- **Les divisions équi-divisantes (ou suppléantes)** permettent de classer un nouveau sujet présentant des données manquantes. L'idée est la suivante : si la meilleure division d* du segment t est obtenue à partir de la variable V_j, on définit pour chaque variable V_i avec i = 1,2,..., p où p est le nombre de variables explicatives et $i \neq j$, la division d_i la plus semblable à d^* : d_i est une division équi-divisante de d*. La meilleure division équi-divisante parmi les (p-1) divisions équi-divisantes d_i ($i \neq j$) est la division équi-divisante la plus semblable à d^*. De la même manière, on peut définir la seconde (meilleure) division équi-divisante, la troisième, etc.

  Remarque : Il est possible que le programme ne fournisse pas de division suppléante bien que la demande d'un certain nombre de ces divisions soit faite dans le fichier de commande. Cette situation peut s'expliquer par l'exemple suivant où, pour une raison de simplicité, les probabilités a priori sont estimées pour les fréquences des groupes dans l'échantillon.

  On considère la meilleure division et la division **triviale** suivantes :

où les nombre indiqués dans les segments sont les effectifs de ces segments. Le tableau de contingence issu du croisement de avec est :

20	10	30
0	0	0
20	10	30

On remarque alors que dans 2/3 des cas cette division prédit correctement . Et donc toute division qui fournirait une prédiction de inférieure à celle fournie par la division ne peut être considérée comme suppléante. C'est le cas, par exemple, de la division suivante :

d_m			
	11	4	15
	9	6	15
	20	10	30

Dans REGAR, la mesure d'association entre d_m et d^* serait négative. Les divisions suppléantes correspondent à des valeurs de cette mesure variant entre 0 (pas d'association) et 1 (association parfaite).

Procédure REGEL : *Elagage de l'arbre pour la régression non paramétrique*

1. Présentation

1.1 Objet

Cette procédure effectue l'édition détaillée d'un sous-arbre de la séquence d'élagage, à partir de la lecture des fichiers NARB et NSEL.

1.2 Editions

Edition du dendogramme de l'arbre, suivi d'une description succincte des divisions. Edition détaillée des segments de l'arbre et écriture de la règle d'affectation aux différents groupes.

1.3 Paramètres

Il y a deux paramètres dans cette procédure : le type d'édition des indivdus dans les segments terminaux (LEDIT) et le nombre de segments terminaux choisis (NOTER). Ce choix détermine l'arbre à éditer.

Attention : ce paramètre doit prendre une des valeurs listées à la fin de l'exécution de REGAR sous le titre "Segments terminaux" (ce sont les seuls arbres qu'il soit possible de construire).

2. Instructions de commande

Les valeurs par défaut sont indiquées entre parenthèses, à la suite du nom du paramètre. Si on désire que les paramètres prennent leurs valeurs par défaut, on codera le mot-clef NOPAR à la place des paramètres repérés par (**3**).

(1) PROC REGEL Edition d'un arbre de segmentation

(2) *titre de la procédure*

(3) NOTER (0) nombre de segments terminaux de l'arbre à éditer.

LEDIT (0 ou NON) édition de la correspondance segments terminaux - individus.

3. Présentation détaillée des paramètres

NOTER **nombre de segments terminaux de l'arbre à éditer**

- *valeurs possibles* : une des valeurs du tableau d'élagage (proc REGAR)
- *valeur par défaut* : 0

La procédure REGAR qui précède a édité le tableau d'élagage. NOTER doit être l'une des valeurs de la colonne "segments terminaux". L'arbre édité aura NOTER segments terminaux.

Si NOTER = 0 et si PRCT ou IVTES est différent de 0 dans REGAR, NOTER correspond à l'arbre repéré par une étoile dans le tableau d'élagage.

Remarque : Il arrive que l'arbre optimum ne comporte qu'un seul segment (l'échantillon n'est pas divisible). Ceci correspond en général à une discrimination difficile à réaliser avec les variables choisies. L'utilisateur peut cependant dans ce cas choisir un arbre sous-optimal dans la liste fournie par REGAR.

LEDIT **édition de la correspondance segments terminaux - individus**

- *valeurs possibles* : 0 ou NON pas d'édition
  1 ou COMPO composition de chaque groupe du nœud terminal
  2 ou AFFEC groupe d'appartenance de chaque individu par nœud terminal
  3 ou TOT options COMPO + AFFEC
- *valeur par défaut* : 0 ou NON

On peut désirer connaître de façon précise la répartition des individus dans les groupes d'origine des nœuds terminaux. Deux types d'éditions sont proposées ici.

- Si LEDIT = COMPO, on obtiendra pour chaque nœud terminal et pour chaque groupe d'origine, la liste des individus qui en font partie.
- Si LEDIT = AFFEC, on donne la liste des individus appartenant au segment terminal, avec pour chacun d'eux le numéro du groupe d'origine auquel il appartient.

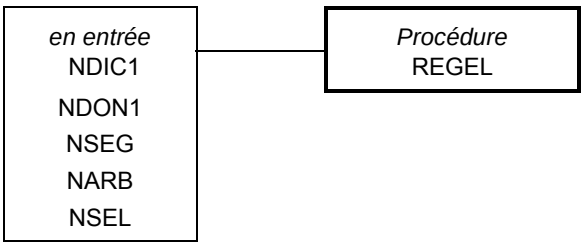
Attention : dans cette sortie, les individus sont repérés par leur identificateur court (en 4 caractères).

4. Exemple

```
----+----1----+----2----+----+----3----+----4----
NDIC1 = 'NDIC1.A' , NDON1 = 'NDON1.A'
NSEG = 'NSEG.A' ,NARB = 'NARB.A' ,NSEL = 'NSEL.A'
LISTP = OUI
PROC REGEL
exemple de procédure regel
NOTER = 0
STOP
----+----1----+----2----+----+----3----+----4----
```

5. Fichiers nécessaires à l'exécution

- en lecture NDIC1 (fichier dictionnaire utile apuré)
  NDON1 (fichier des données utiles apuré)
  NSEG (fichier des données transposées)
  NARB (fichier contenant l'arbre de décision binaire)
  NSEL (fichier des élagages)



Partie 2 :

Commentaires d'exemples

Régression,
Discrimination

Dans cette partie nous présentons l'application des deux méthodes sur deux exemples :

- Estimation d'une variable "survie en mois" (exemple 1).
- Discrimination de la variable "sinistralité" a 2 modalités : 0 sinistre / 1 sinistre et plus (exemple 2).

Exemple 1 : Régression par arbre de décision binaire

Pour cette première application, on dispose de 589 individus, malades atteints d'un cancer. La variable à expliquer est la variable "survie en mois", qui est en moyenne de 23,73 dans l'échantillon total. Les variables explicatives sont toutes nominales. Ce sont des variables décrivant l'apparition des métastases, l'endroit d'apparition, le suivi ou non de chimiothérapie.

```
-------------------------------------------------------------------------------
   1 . chimio                                                (   2 MODALITES )
   2 . delai apparition metastases                           (   4 MODALITES )
   3 . Karnofsky                                             (   2 MODALITES )
   4 . metastases foie                                       (   2 MODALITES )
   5 . metastases poumon                                     (   2 MODALITES )
   6 . metastases plevre                                     (   2 MODALITES )
   7 . metastases os                                         (   2 MODALITES )
   8 . metastases cutanees                                   (   2 MODALITES )
   9 . nb sites metastatiques                                (   3 MODALITES )
  10 . lactico dehydrogenase                                 (   3 MODALITES )
  11 . lymphocytes                                           (   2 MODALITES )
  12 . albumine                                              (   2 MODALITES )
  13 . survie en mois                                 ( CONTINUE )
-------------------------------------------------------------------------------
```

On cherche quelles sont les variables qui ont la plus grande influence sur la survie en mois.

1. Sorties de REGAR

VARIABLE A EXPLIQUER (QUANTITATIVE) = survie en mois

ECHANTILLON TEST TIRE AU HASARD, PTEST = 33.00%

FREQUENCES DES ECHANTILLONS

MOY = MOYENNE DE LA VARIABLE A EXPLIQUER
EC = ECART TYPE DE LA VARIABLE A EXPLIQUER

ECHANTILLON	EFFECTIF	MOY	EC
BASE	385	24.0130	21.0746
TEST	204	23.2010	20.6541

L'échantillon-test (33% de l'échantillon total) est tiré au hasard. On a ici les moyennes et écarts-type de la variable "survie en mois" (variable à expliquer) à la racine de l'arbre, dans l'échantillon de base et dans l'échantillon-test.

Si l'arbre complet (arbre avant élagage) présente un nombre de niveaux inférieur ou égal à 13, il est dessiné. Par contre la description de l'arbre complet sous forme de tableau est fournie quel que soit le nombre de niveaux.
Des explications concernant le dessin de l'arbre et sa description sont données respectivement dans **les sorties S_3** et **S_4**.

S_1

SOUS-ARBRE OPTIMAL (*)

ARBRE	SEGMENTS TERMINAUX	ERREUR RELATIVE DE PREDICTION ECH. TEST	ECH. BASE	OPTIMUM
A1	144	1.2439 +/- .170	.3209	
A2	143	1.2443 +/- .170	.3209	
A3	142	1.2443 +/- .170	.3209	
A4	141	1.2441 +/- .170	.3209	
.	.	.	.	
.	.	.	.	
.	.	.	.	
A91	8	.8739 +/- .063	.7199	
A92	7	.8516 +/- .061	.7351	
A93	5	.8428 +/- .054	.7690	
A94	**4**	**.8064 +/- .045**	**.7941**	*
A95	2	.9087 +/- .027	.8879	
A96	1	1.0000 +/- .006	1.0000	

ERREUR INITIALE (S2) = 444.1375

CONCLUSION

SOUS-ARBRE OPTIMAL N° = 94
NOMBRE DE SEGMENTS TERMINAUX = 4

S_1 fournit l'arbre complet $A_{max} = A_1$ qui contient 144 segments terminaux.

La procédure d'élagage entraîne la suppression des branches les moins informatives de l'arbre et produit cette séquence de sous-arbres (A_1, A_2,...A_{96}). Ce tableau fournit pour chaque sous-arbre de la séquence :

- son nombre de segments terminaux
- l'erreur relative de prédiction qui lui est associée dans l'Echantillon de Base et dans l'Echantillon-Test.
  - Pour l'échantillon de base cette erreur relative représente l'Erreur Apparente de Prédiction (EAP) rapportée à la variance de l'échantillon de base à la racine qui est : $(21,0746)^2 = 444,138$.
  - Pour l'échantillon-test, il s'agit de l'estimation de l'Erreur Théorique de Prédiction rapportée à la variance de l'échantillon-test à la racine de l'arbre ($(20,6541)^2 = 426,592$).

Les éléments du tableau **S_1** permettent d'obtenir le sous-arbre optimal repéré par un * dans la dernière colonne. Il s'agit du plus petit (en terme de nombre de segments terminaux) sous-arbre de la séquence correspondant à la plus petite estimation de l'erreur théorique de prédiction. Dans cet exemple le sous-arbre en question est A_{94} contenant 4 segments terminaux dont l'estimation de l'erreur relative théorique de prédiction est 0,8064.

Remarque :
Dans la procédure REGAR, le sous-arbre sélectionné est obtenu en utilisant la "règle d'un écart-type" : c'est celui dont l'estimation de l'erreur théorique est juste inférieure ou égale à 0,8064 + 0,045 = 0,8514. Avec cette règle c'est toujours A_{94} qui est sélectionné dans cet exemple.

L'**erreur relative** correspondant à l'échantillon de base représente la variance résiduelle de l'arbre, rapportée à la variance s^2 de Y à la racine de l'arbre (ou erreur apparente initiale), c'est-à-dire, la **part de la variance non expliquée par l'arbre de régression** construit à partir de l'échantillon de base.

L'erreur relative correspondant à l'échantillon-test représente l'estimation de la part de la variance théorique non expliquée par l'arbre de régression. C'est cette erreur que l'on doit considérer avant d'utiliser la règle de prédiction fournie par l'arbre.

2. Sorties de REGEL

S2

```
------------------------------------------------------------------------------------
                              DESSIN DE L'ARBRE - LEGENDE
------------------------------------------------------------------------------------
   1.  T        : NUMERO DU SEGMENT
   2.  N(T)     : NOMBRE D'INDIVIDUS CONTENUS DANS T
   3.  M(T)     : MOYENNE PREDITE DANS LE SEGMENT T
   4.  S2(Y/T)  : ERREUR DE PREDICTION DANS LE SEGMENT T

-----------------------------------------------------------------------------------------
         ARBRE ELAGUE (ECHANTILLON DE BASE), ERREUR APPARENTE DE PREDICTION =   352.6807
-----------------------------------------------------------------------------------------
T   N(T)    M(T)    S2(Y/T)                                    SEGM.   COUPURE
2     73    9.42     80.76   .  .  .  .  .  .  .  .   2---+
                                                          |
                                                      1---+    1     VAR 10 = ldh3
                                                          |
6     74    18.54    313.09 .  .  .  .  .  .  .    6---+   |
                                                      |   |
                                                      3---+        3     VAR  3 = karm
                                                      |
14    79    38.61    674.47 .  .  .  .  .    14---+   |
                                              |   |
                                              7---+            7     VAR  2 = dlm1 dlm4
                                              |
15  159     26.01    336.07 .  .  .  .  .  15---+
```

S2 fournit l'arbre de prédiction binaire (l'arbre élagué à partir de l'arbre complet A_{max}) :

- variables explicatives
- règle de prédiction de la valeur de Y d'un nouvel individu.

S'2

```
--------------------------------------------------------------------------------------------
  ARBRE ELAGUE (ECHANTILLON TEST), ESTIMATION DE L'ERREUR THEORIQUE DE PREDICTION = 344.5525
--------------------------------------------------------------------------------------------
 T   N(T)    M(T)  S2(Y/T)                                          SEGM.   COUPURE
 2     34    9.06   73.94  .  .  .  .  .  .  .  .     2---+
                                                        |
                                                    1---+    1     VAR 10  =  ldh3
                                                        |
 6     44   17.43  291.56  .  .  .  .  .  .  .    6---+   |
                                                    |   |
                                                    3---+        3     VAR  3  =  karm
                                                        |
14     35   38.43  679.45  .  .  .  .  .     14---+   |
                                               |   |
                                               7---+            7     VAR  2  =  dlm1  dlm4
                                               |
15     91   25.42  341.47  .  .  .  .  .     15---+
```

S'2 fournit les renseignements concernant le passage des individus de l'échantillon-test dans l'arbre élagué. A cet arbre élagué est associée l'estimation de l'erreur théorique de prédiction.
La règle de prédiction qui est détaillée dans la sortie **S5** ne peut être utilisée que si l'estimation de l'erreur théorique de prédiction est jugée petite compte tenu du problème de régression posé. De plus la prédiction de la valeur de y pour un nouvel individu est d'autant plus précise que la variance du segment terminal auquel il est affecté est petite.

S_3

DESCRIPTION DE L'ARBRE ELAGUE (ECHANTILLON DE BASE)

NOMBRE TOTAL DE SEGMENTS : 7
NOMBRE DE SEGMENTS TERMINAUX : 4

MOY = MOYENNE PREDITE
S2(Y/T) = ERREUR DE PREDICTION
VAR = NUMERO D'ORDRE DE LA VARIABLE DE COUPURE

SEGM.	EFF.	%	MOY	S2(Y/T)	VAR	LIBELLE DE LA VARIABLE DE COUPURE	COUPURE
1	385	100.00	24.013	21.075	10	lactico dehydrogenase	ldh3
2	73	18.96	9.425	80.765		. . . . SEGMENT TERMINAL	
3	312	81.04	27.426	21.627	3	Karnofsky	karm
6	74	19.22	18.541	313.086		. . . . .SEGMENT TERMINAL	
7	238	61.82	30.189	21.991	2	delai apparition metastases	dlm1
							dlm4
14	79	20.52	38.608	674.466		. . . . SEGMENT TERMINAL	
15	159	41.30	26.006	336.069		. . . . .SEGMENT TERMINAL	

Ce tableau contient la description de l'arbre élagué (pour l'échantillon de base) :

- les différentes divisions
- l'effectif (et la proportion) des différents segments intermédiaires ou terminaux.
- la moyenne prédite et l'erreur de prédiction des différents segments.

Ce tableau permet, surtout si on demande l'édition de l'arbre complet, de connaître les effectifs des profils correspondant aux différents chemins de l'arbre.

Ainsi par exemple, le segment n° 14 contient les 79 individus de l'échantillon de base caractérisés par un délai dlm1 ou dlm4, une bonne valeur du Karnofsky et une valeur de ldh comprise dans les classes 1 ou 2. Ces sujets ont une durée moyenne de survie de 38 mois avec un écart-type de 26 mois.

Sachant que les deux fils descendants du segment n° t sont numérotés dans le programme 2t (descendant gauche) et 2t + 1 (descendant droite) il est aisé, en se servant des coupures de reconstituer le chemin parcouru pour tomber dans n'importe quel segment.

Exemple :

Remarque :
La coupure fournie dans la dernière colonne caractérise la branche descendante gauche. Plus précisément, si la coupure est une modalité, les individus présentant cette modalité constituent le segment descendant gauche, et ceux présentant une modalité autre que celle-ci forment le segment descendant droit. Si la coupure est une valeur d'une variable continue, le segment descendant gauche est formé des individus présentant une valeur de cette variable inférieure ou égale à cette coupure.

S'3

PASSAGE DES INDIVIDUS DE L'ECHANTILLON TEST DANS L'ARBRE ELAGUE

NOMBRE TOTAL DE SEGMENTS : 7
NOMBRE DE SEGMENTS TERMINAUX : 4

MOY = MOYENNE PREDITE
S2(Y/T) = ERREUR DE PREDICTION
VAR = NUMERO D'ORDRE DE LA VARIABLE DE COUPURE

SEGM.	EFF.	%	MOY	S2(Y/T)	VAR	LIBELLE DE LA VARIABLE DE COUPURE	COUPURE
1	204	100.00	23.201	20.654	10	lactico dehydrogenase	ldh3
2	34	16.67	9.059	73.938		. . . . . .SEGMENT TERMINAL . . .	
3	170	83.33	26.029	21.193	3	Karnofsky	karm
6	44	21.57	17.432	291.564		. . . . . .SEGMENT TERMINAL	
7	126	61.76	29.032	21.664	2	delai apparition metastases	dlm1
							dlm4
14	35	17.16	38.429	679.445		. . . . . .SEGMENT TERMINAL. . . .	
15	91	44.61	25.418	341.474		. . . . . .SEGMENT TERMINAL	

Ce tableau concerne les mêmes renseignements que ceux fournis par **S3** mais pour le passage des individus de l'échantillon-test dans l'arbre élagué.

S4

DESCRIPTION DES COUPURES
ECHANTILLON BASE

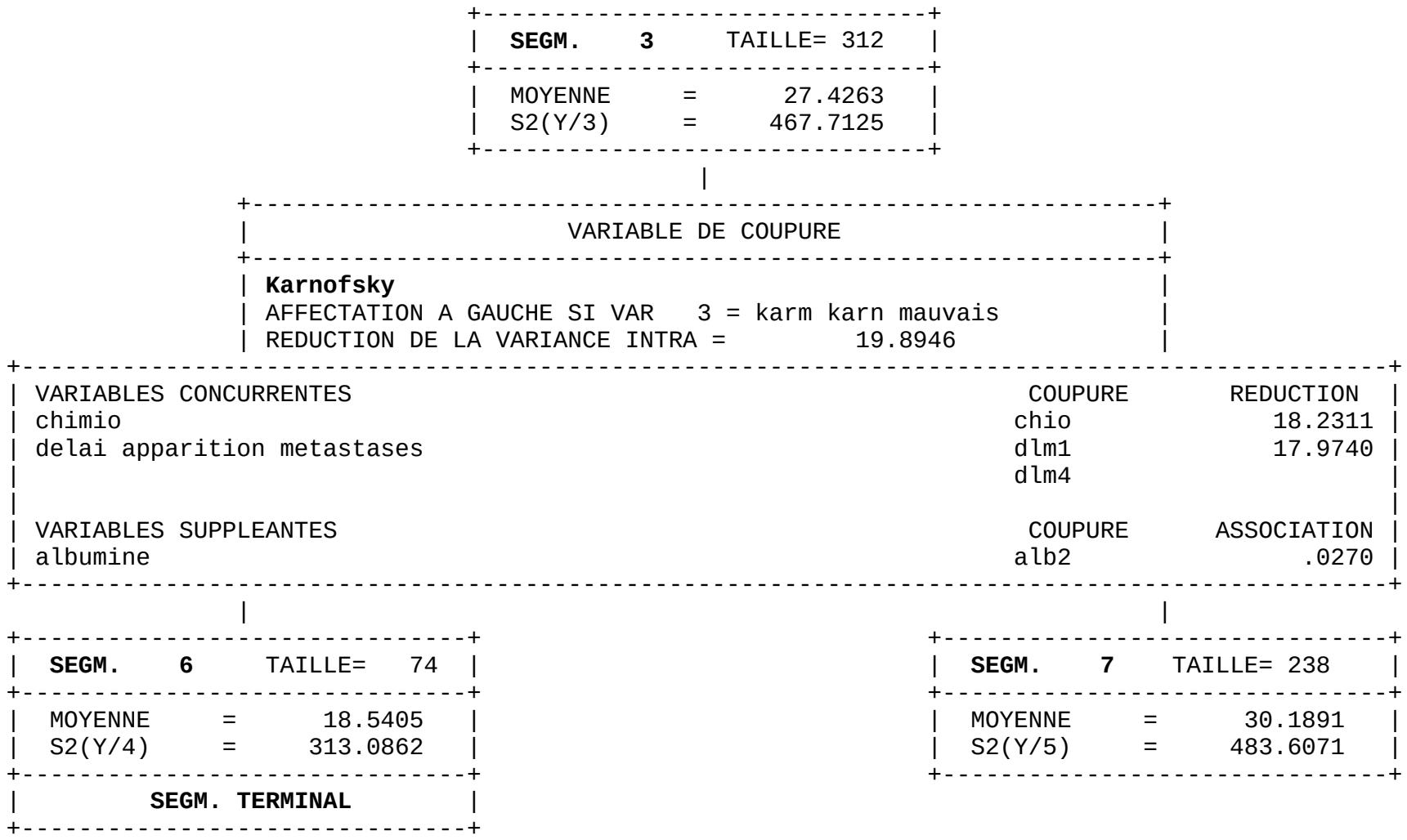
SEGM. 1 TAILLE= 385
MOYENNE = 24.0130
S2(Y/1) = 444.1375

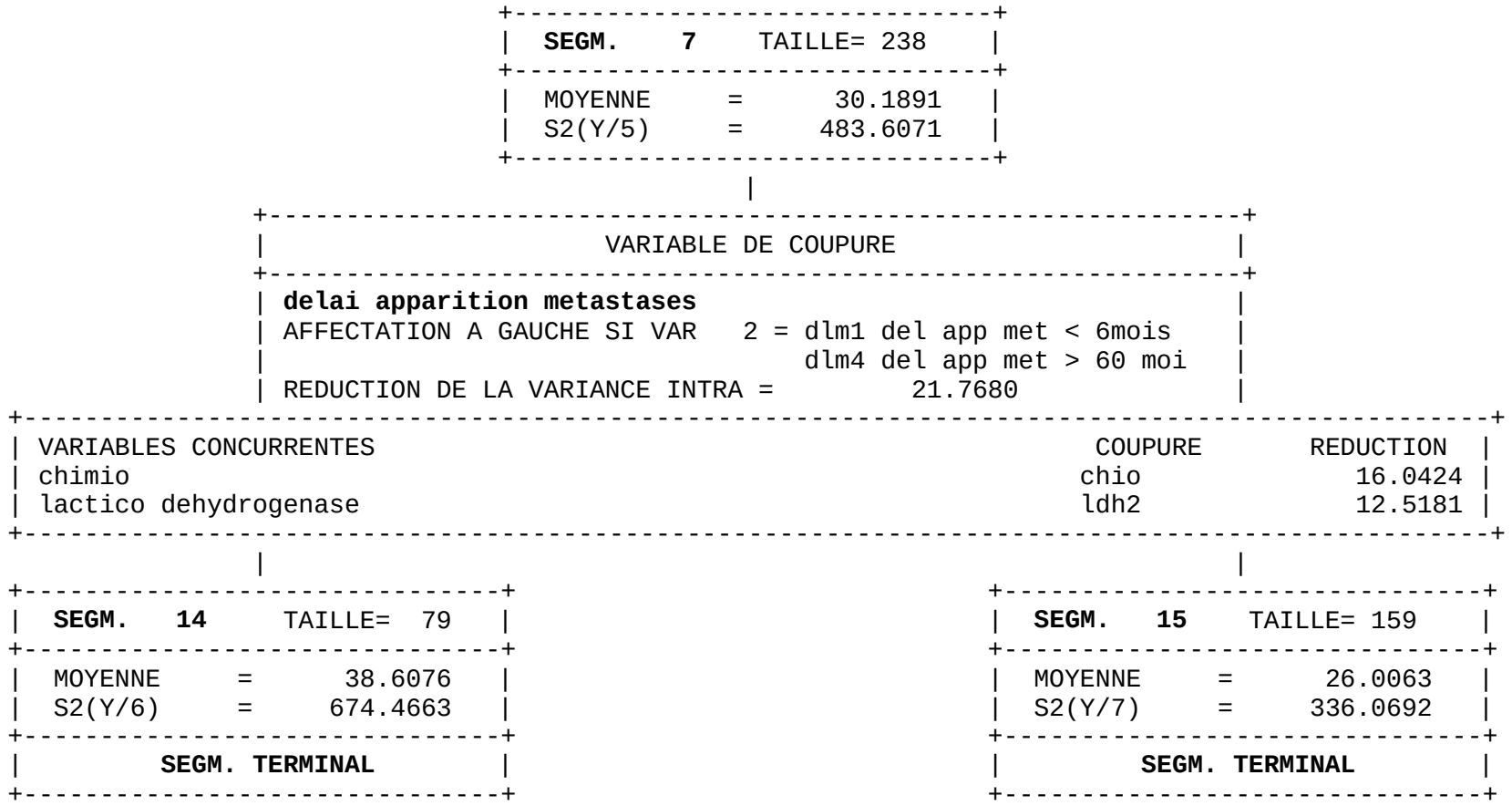
VARIABLE DE COUPURE
lactico dehydrogenase
AFFECTATION A GAUCHE SI VAR 10 = ldh3 ldh (>380)
REDUCTION DE LA VARIANCE INTRA = 49.7943

VARIABLES CONCURRENTES	COUPURE	REDUCTION
metastases foie	foio	26.9570
Karnofsky	karm	26.5046

SEGM. 2 TAILLE= 73
MOYENNE = 9.4247
S2(Y/2) = 80.7649
SEGM. TERMINAL

SEGM. 3 TAILLE= 312
MOYENNE = 27.4263
S2(Y/3) = 467.7125


SEGM. 3 TAILLE= 312
MOYENNE = 27.4263
S2(Y/3) = 467.7125
VARIABLE DE COUPURE
Karnofsky
AFFECTATION A GAUCHE SI VAR 3 = karm karn mauvais
REDUCTION DE LA VARIANCE INTRA = 19.8946
VARIABLES CONCURRENTES
COUPURE
REDUCTION
chimio
chio
18.2311
delai apparition metastases
dlm1
17.9740
dlm4
VARIABLES SUPPLEANTES
COUPURE
ASSOCIATION
albumine
alb2
.0270
SEGM. 6 TAILLE= 74
MOYENNE = 18.5405
S2(Y/4) = 313.0862
SEGM. TERMINAL
SEGM. 7 TAILLE= 238
MOYENNE = 30.1891
S2(Y/5) = 483.6071



La sortie **S4** fournit les informations détaillées de tous les segments intermédiaires et terminaux et de toutes les divisions (ou coupures) de l'arbre retenu :

- pour chaque segment la taille, la moyenne et son erreur de prédiction.
- pour chaque division (ou coupure)
  - la taille du segment parent et sa composition, de même que les tailles et compositions de ses fils descendants.
  - la réduction de l'impureté en passant du segment parent à ses fils descendants. Cette réduction de l'impureté est la plus grande des réductions entraînées par **toutes** les divisions ou coupures admissibles.

Les deux meilleures divisions concurrentes et les deux meilleures divisions suppléantes demandées sont ici décrites. Elles correspondent respectivement aux deux divisions entraînant les plus grandes réductions de la variance intra qui sont inférieures à celle entraînée par la division sélectionnée, et aux deux divisions les plus associées à la division sélectionnée. La mesure d'association entre deux divisions varie de 0 (pas d'association) à 1 (association parfaite).

Remarque :
Un **C** au niveau de la coupure d'une variable suppléante signifie que c'est la branche descendante droite qui est caractérisée par les individus qui présentent cette coupure s'il s'agit d'une modalité, ou une valeur inférieure ou égale à cette coupure s'il s'agit d'une variable continue.

S_5

```
-------------------------------------------------------------------------------------------
                                DESCRIPTION DES SEGMENTS TERMINAUX
                                ECHANTILLON DE BASE, EFF. =    385
-------------------------------------------------------------------------------------------

SEGMENT NUMERO N =     2

   % DE L'ECHANTILLON BASE        =     18.961
   MOYENNE                        =      9.425
   ERREUR APPARENTE DE PREDICTION =     80.765

REGLE D'AFFECTATION
+-----------------------------------------------------------------------------------------+
|  lactico dehydrogenase                                        =   ldh3                  |
+-----------------------------------------------------------------------------------------+


SEGMENT NUMERO N =     6
   % DE L'ECHANTILLON BASE        =     19.221
   MOYENNE                        =     18.541
   ERREUR APPARENTE DE PREDICTION =    313.086
```

```
REGLE D'AFFECTATION
+-------------------------------------------------------------------------------------+
|  lactico dehydrogenase                                    =   ldh1    ldh2          |
|  Karnofsky                                                =   karm                  |
+-------------------------------------------------------------------------------------+
```

SEGMENT NUMERO N = 14

```
   % DE L'ECHANTILLON BASE        =     20.519
   MOYENNE                        =     38.608
   ERREUR APPARENTE DE PREDICTION =    674.466

REGLE D'AFFECTATION
+-------------------------------------------------------------------------------------+
|  lactico dehydrogenase                                    =   ldh1    ldh2          |
|  Karnofsky                                                =   karn                  |
|  delai apparition metastases                              =   dlm1    dlm4          |
+-------------------------------------------------------------------------------------+
```

SEGMENT NUMERO N = 15

```
   % DE L'ECHANTILLON BASE        =     41.299
   MOYENNE                        =     26.006
   ERREUR APPARENTE DE PREDICTION =    336.069

REGLE D'AFFECTATION
+-------------------------------------------------------------------------------------+
|  lactico dehydrogenase                                    =   ldh1    ldh2          |
|  Karnofsky                                                =   karn                  |
|  delai apparition metastases                              =   dlm2    dlm3          |
+-------------------------------------------------------------------------------------+
```

La sortie **S5** donne pour chacun des segments terminaux le pourcentage d'individus de l'échantillon de base affectés au segment, la moyenne prédite de Y ainsi que l'erreur apparente de prédiction qui lui est associée.

S'5

```
--------------------------------------------------------------------------------
                          ESTIMATION DE L'ERREUR THEORIQUE DE PREDICTION (ETP)
                                    DANS LES SEGMENTS TERMINAUX
--------------------------------------------------------------------------------

SEGMENT NUMERO N =     2

   % DE L'ECHANTILLON TEST =     16.667
   ESTIMATION DE L'ETP     =     73.938


SEGMENT NUMERO N =     6

   % DE L'ECHANTILLON TEST =     21.569
   ESTIMATION DE L'ETP     =    291.564


SEGMENT NUMERO N =    14

   % DE L'ECHANTILLON TEST =     17.157
   ESTIMATION DE L'ETP     =    679.445


SEGMENT NUMERO N =    15

   % DE L'ECHANTILLON TEST =     44.608
   ESTIMATION DE L'ETP     =    341.474
```

S'5 fournit pour chacun des segments terminaux le pourcentage d'individus de l'échantillon-test affectés au segment, ainsi que l'estimation de l'erreur théorique de prédiction.

En résumé :
dans cet exemple, le programme de régression fournit un arbre de prédiction à 4 nœuds terminaux, dont l'estimation de l'erreur relative de prédiction est égale à 0,81 ; ce qui correspond, par analogie à la régression multiple, à un coefficient de corrélation multiple R^2 = 1 - 0,81 = 0,19 qui représente le pourcentage de la variance expliquée par l'arbre.

La règle de prédiction fournie est la suivante :
Si un nouvel individu présente une valeur de ldh comprise dans l'intervalle ldh3, la valeur prédite de Y est égale à 9,425 et l'erreur de prédiction commise est de $(8,599)^2$. Et ainsi de suite pour les différents segments terminaux.

Remarque concernant les individus anonymes
Pour les individus anonymes, reprérés par une valeur manquante pour la variable à expliquer, le programme fournit leur identificateur, le segment terminal auquel ils appartiennent, la valeur prédite de Y, et l'erreur de prédiction associée.

Exemple 2 : Discrimination par arbre de décision binaire

Cet exemple présente des données sur des assurés d'assurances belges. Ces dernières cherchent à savoir ce qui discrimine les individus ayant eu 0 sinistre durant la dernière année, des individus ayant eu 1 sinistre ou plus. On dispose d'un échantillon de 1106 assurés, dont environ la moitié a déclaré aucun sinistre (556), et l'autre moitié (550) un sinistre et plus. Les variables explicatives, toutes nominales sont :

Variables explicatives

2	Code usage - CUSA 5-6	(2 MODALITES)
4	Sexe - SEXE 11-12	(3 MODALITES)
5	Code Langue - CLAN 14-15	(2 MODALITES)
14	Présence garantie 2 - GAR2 41-42	(2 MODALITES)
15	Presence garantie 3 - GAR3 44-45	(2 MODALITES)
16	Primes Acquises RC 1991 en francs belges - TPA11	(CONTINUE)
24	Age de l'assuré (3 mod) - DNAI 8-9	(3 MODALITES)
25	Code postal souscripteur (2 mod) - POSS2 17-18	(2 MODALITES)
26	Bonus-malus Année -1 (2 mod) - GBM1	(2 MODALITES)
27	Date effet Police (2 mod) - DPEP 26-27	(2 MODALITES)
28	Puissance du véhicule (2 mod) - PUIS 32-33	(2 MODALITES)
29	Année de construction du véhicule (2 mod) - DCOS 38-39	(2 MODALITES)
30	Primes acquises RC 1991 en francs belges (2 mod) - TP11	(3 MODALITES)

Variable à discriminer

1	Sinistralité RC - SNB11 2-3	(2 MODALITES)

1. Sorties de DISAR

VARIABLE DE GROUPE = Sinistralité RC - SNB11 2-3

ECHANTILLON TEST TIRE AU HASARD, PTEST = 33.00%

PROBABILITE A PRIORI ET FREQUENCES DES ECHANTILLONS

	PROBA	ECHANTILLON DE BASE		ECHANTILLON TEST	
CLASSE	A PRIORI	EFF.	%	EFF.	%
SRC0	.503	373	50.34	183	50.14
SRC1	.497	368	49.66	182	49.86
TOTAL	1.000	741	100.00	365	100.00

COUTS DE MAUVAIS CLASSEMENT

CLASSE D'AFFECTATION	CLASSE D'ORIGINE SRC0	SRC1
SRC0	.00	1.00
SRC1	1.00	.00

Dans cet exemple les probabilités a priori sont prises égales aux fréquences des deux groupes dans l'échantillon. Un échantillon-test est tiré au hasard et les coûts de mauvais classement sont égaux à 1
Si l'arbre complet (arbre avant élagage) présente un nombre de niveaux inférieur ou égal à 13, il est dessiné. Par contre la description de l'arbre complet sous forme de tableau est fournie quel que soit le nombre de niveaux. Des explications concernant le dessin de l'arbre et sa description sont données respectivement dans les impressions **S_3** et **S_4**.

S_1

SEQUENCE D'ELAGAGE

ARBRE	SEGMENTS TERMINAUX												
A1 (29 ST)	16	17	36	37	19	10	22	368	738	739	185	93	94
	95	12	13	56	57	58	118	238	478	958	1918	3838	7678
	15358	15359	15										
A2 (18 ST)	16	17	9	10	22	184	185	93	47	12	13	28	58
	118	238	478	479	15								
A3 (13 ST)	16	17	9	5	12	13	28	58	118	238	478	479	15
A4 (7 ST)	16	17	9	5	12	13	7						
A5 (4 ST)	2	12	13	7									
A6 (2 ST)	2	3											
A7 (1 ST)	1												

L'arbre A_{max} (noté A_1) du tableau **S_1** contient 29 nœuds terminaux. Le critère d'élagage entraîne la suppression de certaines branches (les moins informatives), et le premier sous-arbre obtenu après élagage (noté A_2) contient 18 nœuds terminaux. Ainsi de suite jusqu'à l'obtention de A_6 qui contient les deux nœuds terminaux (2 et 3) et de A_7 (racine de l'arbre).

S_2

SOUS-ARBRE OPTIMAL (*)

ARBRE	SEGMENTS TERMINAUX	COUT RELATIF ECH. TEST	COUT RELATIF ECH. BASE	OPTIMUM
A1	29	.2967 +/- .037	.2582	
A2	18	.2857 +/- .037	.2582	
A3	13	.2857 +/- .037	.2609	
A4	7	.2857 +/- .037	.2663	
A5	**4**	**.3132 +/- .038**	**.2935**	*
A6	2	.3571 +/- .040	.3207	
A7	1	1.0000 +/- .052	1.0000	

COUT D'ERREUR INITIAL = .4966
AFFECTATION INITIALE = SRC0

CONCLUSION

SOUS-ARBRE OPTIMAL N° = 5
NOMBRE DE SEGMENTS TERMINAUX = 4

VARIABLE DE GROUPE = Sinistralité RC - SNB11 2-3

DESCRIPTION DE L'ARBRE APRES ELAGAGE
ARBRE CHOISI : 4 SEGMENTS TERMINAUX

ECHANTILLON	%	EFFECTIF
BASE	67.00	741
TEST	33.00	365
TOTAL	100.00	1106

Le tableau de la sortie **S2** fournit pour chaque sous-arbre de la séquence (A_1 , ..., A_7) obtenue :

- son nombre de segments terminaux
- le coût **relatif** qui lui est associé pour l'échantillon de base et pour l'échantillon-test.

- Pour l'échantillon de base ce coût relatif est le coût d'erreur apparent rapporté au coût d'erreur initial, soit 49,66 % (voir page Commentaires 19).
- Pour l'échantillon-test, il s'agit de l'estimation du coût d'erreur théorique rapportée à l'estimation de ce taux à la racine de l'arbre soit 49,86 %. A cette estimation du coût d'erreur théorique est associé son écart-type.

Remarque : Les coûts d'erreur apparent et théorique correspondent, dans cet exemple, aux TEA (taux d'erreur apparent) et TET (taux d'erreur théorique) puisque les probabilités a priori sont prises égales aux fréquences des groupes dans l'échantillon, et les coûts de mauvais classement sont tous égaux à 1.

Exemple :
Arbre A_5 à 4 segments terminaux

associé à

écart-type

On en déduit l'écart-type du coût relatif associé à A_5 en rapportant cet écart-type au de la racine, soit .

Les éléments de la sortie **S2** permettent d'obtenir le sous-arbre optimal repéré par une * dans la dernière colonne. Il s'agit du **plus petit** sous-arbre en terme de nombre de segments terminaux de la séquence correspondant au plus petit coût relatif obtenu à l'aide de l'échantillon-test.
Dans cet exemple, le sous-arbre ayant le plus petit coût relatif obtenu à l'aide de l'échantillon-test est A_4 contenant 7 segments terminaux dont le coût relatif de l'échantillon-test est 0,2857. Mais l'arbre sélectionné est l'arbre obtenu en utilisant la "règle d'un écart-type" : c'est celui dont le coût relatif de l'échantillon-test est juste inférieur ou égal à 0,2857 + 0,037 = 0,3227. Il s'agit de l'arbre A_5 à 4 segments terminaux dont le coût relatif est plus grand que 0,2857 mais dont le nombre de segments terminaux est plus petit.

Remarque : Le coût relatif permet d'avoir une idée précise sur la performance de la règle discriminante.

2. Sorties de DISEL

S_3

```
-------------------------------------------------------------------------------------------------------------
                                        DESSIN DE L'ARBRE - LEGENDE
-------------------------------------------------------------------------------------------------------------

   1.     T      NUMERO DU SEGMENT.
   2.    '*'     CLASSE D'AFFECTATION DU SEGMENT T.
                 CLA(T)='*'.
   3.   R(*/T)   RISQUE (OU COUT MOYEN) D'ERREUR ENTRAINE
                 PAR L'AFFECTATION DU SEGMENT T A LA CLASSE '*'.
   4.    P(T)    PROBABILITE D'ARRIVER DANS LE SEGMENT T.


-------------------------------------------------------------------------------------------------------------
                          ARBRE ELAGUE (ECHANTILLON DE BASE), COUT APPARENT =     .1457
-------------------------------------------------------------------------------------------------------------

    T  '*' R(*/T)  P(T)                                                          NO   COUPURE
    2  SRC1  .17   .51  . . .. . . . . . . . . . . . . . . . . .     2---+
                                                                        |
                                                                    1---+    1   VAR 26    =  BM02
                                                                        |
   12  SRC0  .37   .06  . . .. . . . . . . . . . . . . . . . .     12---+      |
                                                                     |      |
                                                                    6---+   |        6   VAR 24    =  AGE1  AGE?
                                                                    |   |   |
   13  SRC1  .19   .02  . . .. . . . . . . . . . . . . . . . .     13---+   |   |
                                                                         |   |
                                                                        3---+        3   VAR 29    =  DC02
                                                                            |
    7  SRC0  .08   .40  . . .. . . . . . . . . . . . . . . . .     7---+
```

S_3 fournit l'arbre de décision binaire élagué: variables discriminantes, règle d'affectation d'un nouvel individu.

Cette règle ne peut être utilisée que si l'estimation du coût d'erreur théorique est jugée assez petite compte tenu du problème de discrimination posé.
L'information fournie, par exemple, pour le segment terminal n° 2 est la suivante : dans l'échantillon de base, la probabilité pour qu'un individu tombe dans le segment n° 2 est de 0,51. S'il y tombe, il est affecté su groupe SCR1 et entraîne un risque d'erreur égal à 0,17.

Remarque:
Le risque d'erreur est égal au pourcentage d'individus mal classés du segment terminal n°2 fourni dans la sortie **S5**, puisque les coûts d'erreur de classement ont été supposés égaux et que les probabilités a priori ont été prises égales aux fréquences des groupes dans l'échantillon.

S'3

```
------------------------------------------------------------------------------------------------------------
                          ARBRE ELAGUE (ECHANTILLON TEST), ESTIMATION DU COUT THEORIQUE =     .1562
------------------------------------------------------------------------------------------------------------

 T  '*' R(*/T)  P(T)                                                                  NO   COUPURE
 2  SRC1  .17   .48  . . . . . . . . . . . . . . . . . . . . . .     2---+
                                                                        |
                                                                    1---+    1   VAR 26    =  BM02
                                                                        |
12  SRC0  .39   .06  . . . . . . . . . . . . . . . . . . . . .    12---+       |
                                                                   |         |
                                                                   6---+     |        6   VAR 24    =  AGE1 AGE?
                                                                   |   |     |
13  SRC1  .10   .03  . . . . . . . . . . . . . . . . . . . . .    13---+     |     |
                                                                   |   |     |
                                                                   3---+        3   VAR 29    =  DC02
                                                                   |
 7  SRC0  .11   .42  . . . . . . . . . . . . . . . . . . . . .     7---+
```

S'3 contient les renseignements des individus de l'échantillon-test passés dans l'arbre élagué avec notamment, le risque global d'erreur estimé à l'aide de l'échantillon-test (estimation du coût théorique = 0,1562) et une estimation du risque d'erreur associée à chaque segment terminal.

S4

```
--------------------------------------------------------------------------------------------------------------
----------------
                                          DESCRIPTION DE L'ARBRE - LEGENDE
--------------------------------------------------------------------------------------------------------------
   1.    T     NUMERO DU SEGMENT.
   2.    N     EFFECTIF DU SEGMENT T.
   3.    %     POURCENTAGE DU SEGMENT T.
   4.   '*'    CLASSE D'AFFECTATION DU SEGMENT T.
   5.  R(*/T)  RISQUE (OU COUT MOYEN) D'ERREUR ENTRAINE
               PAR L'AFFECTATION DU SEGMENT T A LA CLASSE '*'.
   6.   P(T)   PROBABILITE D'ARRIVER DANS LE SEGMENT T.
   7.   VAR    NUMERO D'ORDRE DE LA VARIABLE DE COUPURE.


--------------------------------------------------------------------------------------------------------------
                                    DESCRIPTION DE L'ARBRE ELAGUE (ECHANTILLON DE BASE)
--------------------------------------------------------------------------------------------------------------
   NOMBRE TOTAL DE SEGMENTS     :     7
   NOMBRE DE SEGMENTS TERMINAUX :     4
+-----------------------------------------------------------------------------------------------------------+
|  T  |  N   |   %   | '*'  | R(*/T) |  P(T)  | VAR | LIBELLE DE LA VARIABLE DE COUPURE                  | COUPURE |
+-----------------------------------------------------------------------------------------------------------+
|   1 |  741 |100.00 | SRC0 |  .50   |  1.00  |  26 | Bonus-malus Année -1 (2 mod) - GBM1                |   BM02  |
|   2 |  380 | 51.28 | SRC1 |  .17   |   .51  |     | . . . . . SEGMENT TERMINAL . . . . . .             |         |
|   3 |  361 | 48.72 | SRC0 |  .15   |   .49  |  29 | Année de construction du véhicule (2 mod) - DCOS 38- |   DC02  |
|   6 |   62 |  8.37 | SRC0 |  .48   |   .08  |  24 | Age de l'assuré (3 mod) - DNAI 8-9                 |   AGE1  |
|     |      |       |      |        |        |     |                                                    |   AGE?  |
|  12 |   46 |  6.21 | SRC0 |  .37   |   .06  |     | . . . . . SEGMENT TERMINAL . . . . . .             |         |
|  13 |   16 |  2.16 | SRC1 |  .19   |   .02  |     | . . . . . SEGMENT TERMINAL . . . . . .             |         |
|   7 |  299 | 40.35 | SRC0 |  .08   |   .40  |     | . . . . . SEGMENT TERMINAL . . . . . .             |         |
+-----------------------------------------------------------------------------------------------------------+
```

Le tableau de la sortie **S4** fournit une information complémentaire surtout si on demande l'édition de l'arbre complet. Il permet alors de connaître les effectifs des profils correspondant aux différents chemins de l'arbre.

Ainsi par exemple, le segment n° 12 contient les 46 individus de l'échantillon dont l'âge est AGE 1 ou AGE ? qui possèdent un véhicule dont l'année de construction est DC02 et qui ont un Bonus-Malus BM01.

Sachant que les deux fils descendants du segment n° t sont numérotés dans le programme 2t (descendant gauche) et 2t + 1 (descendant droite), il est aisé, en se servant des coupures, de reconstituer le chemin parcouru pour tomber dans n'importe quel segment.

Exemples:

Remarque :
La coupure fournie dans la dernière colonne caractérise la branche descendante gauche. Plus précisément, si la coupure est une modalité, les individus présentant cette modalité constituent le segment descendant gauche, et ceux présentant une modalité autre que celle-ci forment le segment descendant droit. Si la coupure est une valeur d'une variable continue, le segment descendant gauche est formé des individus présentant une valeur de cette variable inférieure ou égale à cette coupure.

S'4

PASSAGE DES INDIVIDUS DE L'ECHANTILLON TEST DANS L'ARBRE ELAGUE

NOMBRE TOTAL DE SEGMENTS : 7
NOMBRE DE SEGMENTS TERMINAUX : 4

T	N	%	'*'	R(*/T)	P(T)	VAR	LIBELLE DE LA VARIABLE DE COUPURE	COUPURE
1	365	100.00	SRC0	.50	1.00	26	Bonus-malus Année -1 (2 mod) - GBM1	BM02
2	177	48.49	SRC1	.17	.48		. . . . . SEGMENT TERMINAL	
3	188	51.51	SRC0	.19	.52	29	Année de construction du véhicule (2 mod) - DCOS 38-	DC02
6	33	9.04	SRC0	.55	.09	24	Age de l'assuré (3 mod) - DNAI 8-9	AGE1
								AGE?
12	23	6.30	SRC0	.39	.06		. . . . . SEGMENT TERMINAL	
13	10	2.74	SRC1	.10	.03		. . . . . SEGMENT TERMINAL	
7	155	42.47	SRC0	.11	.42		. . . . . SEGMENT TERMINAL	

S'4 fournit des renseignements analogues à ceux de **S4** mais qui concernent le passage des individus de l'échantillon-test dans l'arbre élagué.

S5

DESCRIPTION DES COUPURES : ECHANTILLON BASE

SEGM. 1 TAILLE = 741

* SRC0 373
SRC1 368

VARIABLE DE COUPURE

Bonus-malus Année -1 (2 mod) - GBM1
AFFECTATION A GAUCHE SI VAR 26 = BM02 Autres B-M (-1)
REDUCTION DE L'IMPURETE = .2325

VARIABLES CONCURRENTES	COUPURE		REDUCTION
Primes Acquises RC 1991 en francs belges - TPA11	14270.000		.1241
Primes acquises RC 1991 en francs belges (2 mod) - TP11	18C		.1039
VARIABLES SUPPLEANTES	**COUPURE**		**ASSOCIATION**
Primes Acquises RC 1991 en francs belges - TPA11	14075.500	C	.5263
Primes acquises RC 1991 en francs belges (2 mod) - TP11	18C		.4654

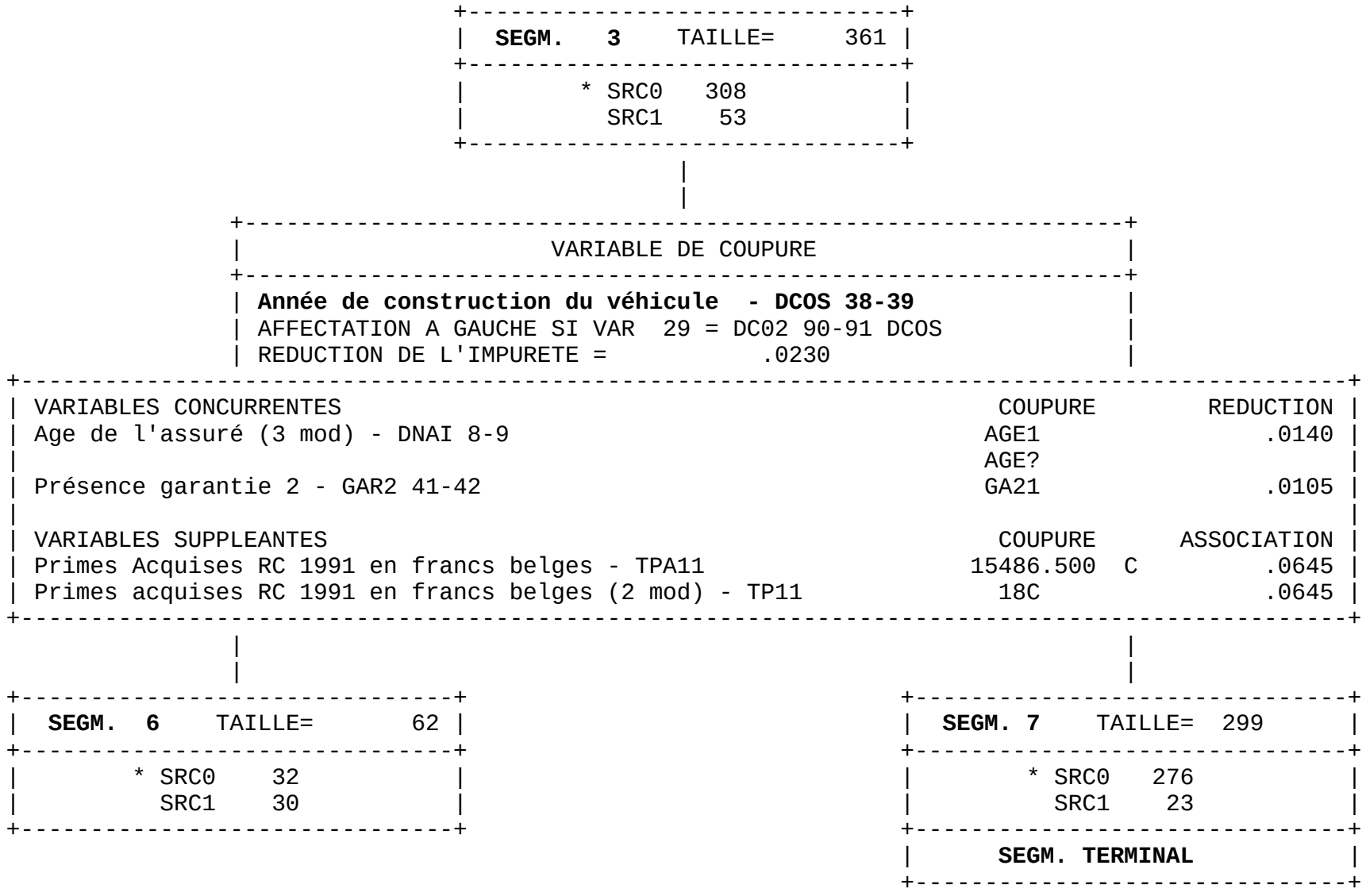
SEGM. 2 TAILLE= 380

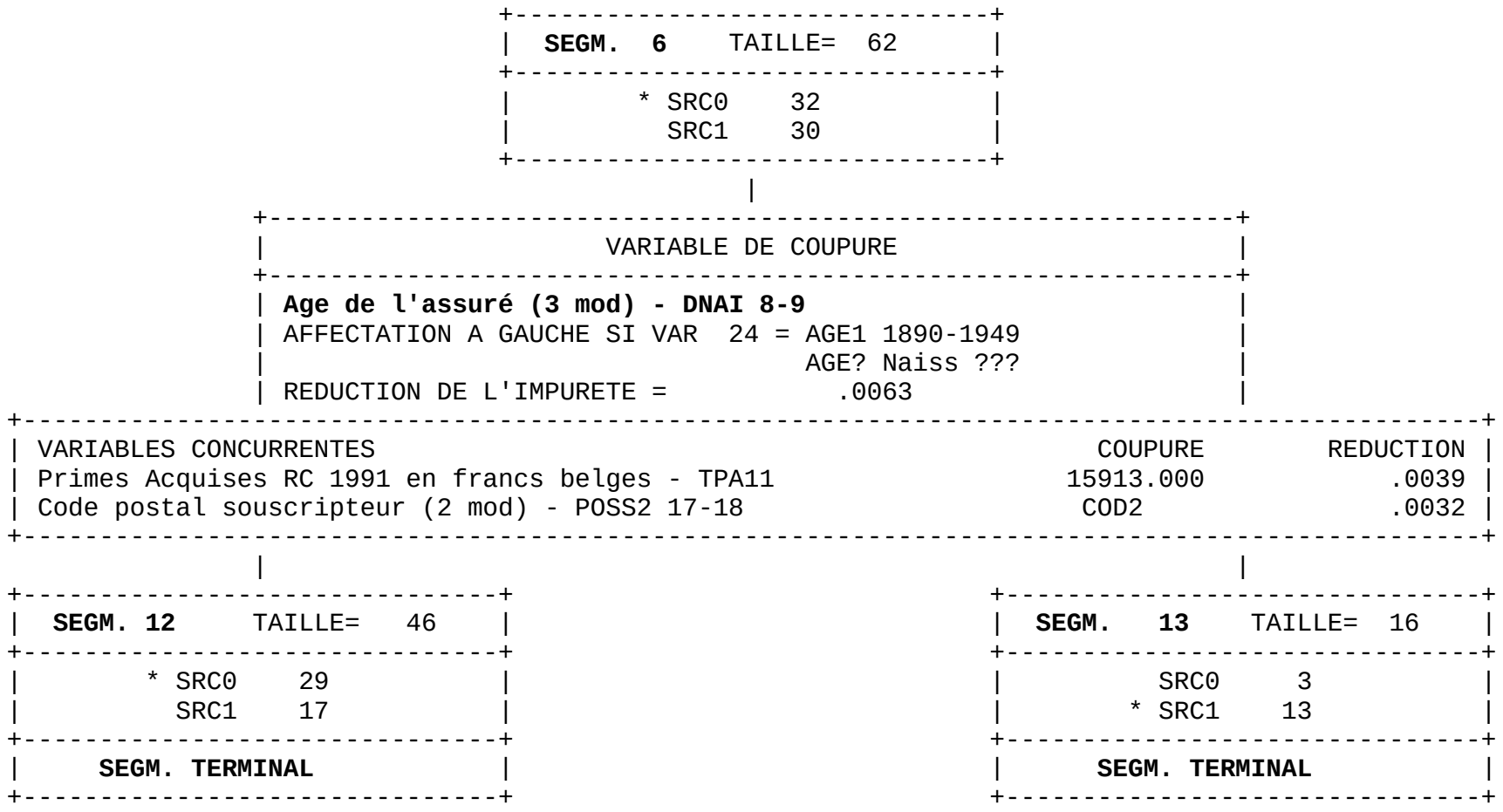
SRC0 65
* SRC1 315

SEGM. TERMINAL

SEGM. 3 TAILLE= 361

* SRC0 308
SRC1 53


SEGM. 3 TAILLE= 361
* SRC0 308
SRC1 53
VARIABLE DE COUPURE
Année de construction du véhicule - DCOS 38-39
AFFECTATION A GAUCHE SI VAR 29 = DC02 90-91 DCOS
REDUCTION DE L'IMPURETE = .0230
VARIABLES CONCURRENTES
COUPURE
REDUCTION
Age de l'assuré (3 mod) - DNAI 8-9
AGE1
.0140
AGE?
Présence garantie 2 - GAR2 41-42
GA21
.0105
VARIABLES SUPPLEANTES
COUPURE
ASSOCIATION
Primes Acquises RC 1991 en francs belges - TPA11
15486.500 C
.0645
Primes acquises RC 1991 en francs belges (2 mod) - TP11
18C
.0645
SEGM. 6 TAILLE= 62
* SRC0 32
SRC1 30
SEGM. 7 TAILLE= 299
* SRC0 276
SRC1 23
SEGM. TERMINAL



L'impression **S5** fournit les informations détaillées de tous les segments (intermédiaires et terminaux) et de toutes les divisions de l'arbre retenu :

- la taille du segment parent et sa composition, de même que les tailles et compositions de ses fils descendants. (*) indique le groupe auquel est affecté le segment.
- la réduction de l'impureté en passant du segment parent à ses fils descendants. Cette réduction de l'impureté est la plus grande des réductions entraînées par **toutes** les divisions ou coupures admissibles.

Les deux meilleures divisions concurrentes et les deux meilleures divisions suppléantes demandées sont ici décrites. Elles correspondent respectivement aux deux divisions entraînant les plus grandes réductions de l'impureté qui sont inférieures à celle entraînée par la division sélectionnée, et aux deux divisions les plus associées à la division sélectionnée. La mesure d'association entre deux divisions varie de 0 (pas d'association) à 1 (association parfaite).

Remarque :
Un **C** au niveau de la coupure d'une variable suppléante signifie que c'est la branche descendante droite qui est caractérisée par les individus qui présentent cette coupure s'il s'agit d'une modalité, ou une valeur inférieure ou égale à cette coupure s'il s'agit d'une variable continue.

S_6

DESCRIPTION DES SEGMENTS TERMINAUX
ECHANTILLON DE BASE, EFF. = 741

LEGENDE
SEGMENT NUMERO = T P(T) : PROBABILITE D'APPARTENIR AU SEGMENT T
(1) CLASSE J (AFFECTATION A J=*)
(2) EFFECTIF DU SEGMENT T DANS LA CLASSE J
(3) EFFECTIF TOTAL DANS LA CLASSE J
(4) PROPORTION DE SUJETS DE LA CLASSE J DANS SEGMENT T
(5) PROPORTION DE SUJETS DU SEGMENT T APPARTENANT A LA CLASSE J
(6) R(J/T) RISQUE D'ERREUR MOYEN ENTRAINE PAR L'AFFECTATION DU SEGMENT T A LA CLASSE J

SEGMENT NUMERO N° = 2 P(2) = .5128

(1) CLASSE J	(2) EFFECT	(3) TOTAL	(4) % T DANS J	(5) % J DANS T	(6) R(J/T)
SRC0	65	373	17.43	17.11	.829
* SRC1	315	368	85.60	82.89	.171
TOTAL	380	741		100.00	

REGLE D'AFFECTATION

Bonus-malus Année -1 (2 mod) - GBM1 = BM02

SEGMENT NUMERO N° = 12 P(12) = .0621

(1) CLASSE J	(2) EFFECT	(3) TOTAL	(4) % T DANS J	(5) % J DANS T	(6) R(J/T)
* SRC0	29	373	7.77	63.04	.370
SRC1	17	368	4.62	36.96	.630
TOTAL	46	741		100.00	

REGLE D'AFFECTATION

Bonus-malus Année -1 (2 mod) - GBM1 = BM01
Année de construction du véhicule (2 mod) - DCOS 38-39 = DC02
Age de l'assuré (3 mod) - DNAI 8-9 = AGE1 AGE?

SEGMENT NUMERO N° = 13 P(13) = .0216

(1) CLASSE J	(2) EFFECT	(3) TOTAL	(4) % T DANS J	(5) % J DANS T	(6) R(J/T)
SRC0	3	373	.80	18.75	.813
* SRC1	13	368	3.53	81.25	.187
TOTAL	16	741		100.00	

```
REGLE D'AFFECTATION
+-------------------------------------------------------------------------------+
|  Bonus-malus Année -1 (2 mod) - GBM1                          =   BM01        |
|  Année de construction du véhicule (2 mod) - DCOS 38-39       =   DC02        |
|  Age de l'assuré (3 mod) - DNAI 8-9                           =   AGE2        |
+-------------------------------------------------------------------------------+
```

SEGMENT NUMERO N° = 7 P(7) = .4035

(1) CLASSE J	(2) EFFECT	(3) TOTAL	(4) % T DANS J	(5) % J DANS T	(6) R(J/T)
* SRC0	276	373	73.99	92.31	.077
SRC1	23	368	6.25	7.69	.923
TOTAL	299	741		100.00	

```
REGLE D'AFFECTATION
+-------------------------------------------------------------------------------+
|  Bonus-malus Année -1 (2 mod) - GBM1                          =   BM01        |
|  Année de construction du véhicule (2 mod) - DCOS 38-39       =   DC01        |
+-------------------------------------------------------------------------------+
```

S6 fournit essentiellement pour chaque segment terminal :

- son groupe d'affectation (* dans la colonne 1) et le risque d'erreur entraîné par l'affectation d'un individu qui tombe dans ce segment terminal (colonne 6).

- la règle d'affectation, c'est à dire plus exactement le chemin à parcourir dans l'arbre pour tomber dans ce segment terminal. Chaque arête de ce chemin correspond à une sélection suivant une ou plusieurs modalités d'une variable si elle est nominale, et, à un intervalle de valeurs d'une variable si elle est ordinale ou quantitative.

S'6

DESCRIPTION DES SEGMENTS TERMINAUX
ECHANTILLON TEST, EFF. = 365

SEGMENT NUMERO N° = 2 P(2) = .4849

(1) CLASSE J	(2) EFFECT	(3) TOTAL	(4) % T DANS J	(5) % J DANS T	(6) R(J/T)
SRC0	30	183	16.39	16.95	.831
* SRC1	147	182	80.77	83.05	.169
TOTAL	177	365		100.00	

SEGMENT NUMERO N° = 12 P(12) = .0630

(1) CLASSE J	(2) EFFECT	(3) TOTAL	(4) % T DANS J	(5) % J DANS T	(6) R(J/T)
* SRC0	14	183	7.65	60.87	.391
SRC1	9	182	4.95	39.13	.609
TOTAL	23	365		100.00	

SEGMENT NUMERO N° = 13 P(13) = .0274

(1) CLASSE J	(2) EFFECT	(3) TOTAL	(4) % T DANS J	(5) % J DANS T	(6) R(J/T)
SRC0	1	183	.55	10.00	.900
* SRC1	9	182	4.95	90.00	.100
TOTAL	10	365		100.00	

SEGMENT NUMERO N° = 7 P(7) = .4247

(1) CLASSE J	(2) EFFECT	(3) TOTAL	(4) % T DANS J	(5) % J DANS T	(6) R(J/T)
* SRC0	138	183	75.41	89.03	.110
SRC1	17	182	9.34	10.97	.890
TOTAL	155	365		100.00	

S'6 fournit la description des segments terminaux équivalente à **S6** mais correspondant aux individus de l'échantillon-test.

S7

AFFECTATION DES INDIVIDUS - RECAPITULATIF

ECHANTILLON BASE = 741

CLASSE D'AFFECTATION	CLASSE D'ORIGINE SRC0	SRC1
SRC0	305	40
	81.8	10.9
SRC1	68	328
	18.2	89.1
TOTAL	373	368
	100.0	100.0

ECHANTILLON TEST = 365

CLASSE D'AFFECTATION	CLASSE D'ORIGINE SRC0	SRC1
SRC0	152	26
	83.1	14.3
SRC1	31	156
	16.9	85.7
TOTAL	183	182
	100.0	100.0

POURCENTAGE APPARENT DE BIEN CLASSES

CLASSE	EFF.	/SUR	%
SRC0	305	373	81.77
SRC1	328	368	89.13
	633	741	85.43

ESTIMATION DU POURCENTAGE THEORIQUE DE BIEN CLASSES

CLASSE	EFF.	/SUR	%
SRC0	152	183	83.06
SRC1	156	182	85.71
	308	365	84.38

S7 fournit les tables de classification (origine, affectation) correspondant à l'échantillon de base et à l'échantillon-test avec, en plus, l'impression du pourcentage apparent de bien classés (échantillon de base) et de l'estimation du pourcentage théorique de bien classés (échantillon-test) **uniquement** si les probabilités a priori sont estimées par les fréquences des groupes dans l'échantillon et si les coûts de mauvais classement sont égaux à 1.

En résumé :
dans cet exemple, le programme de discrimination par l'arbre fournit un arbre de décision à 4 nœuds terminaux, dont l'estimation du coût **relatif** est égale à 0,31, ce qui entraîne une estimation du pourcentage de bien classés égal à 84 %.
La règle d'affectation fournie est la suivante :
Si un individu présente pour la variable Bonus-Malus la modalité BMO2, il sera affecté au groupe SCRI avec un coût d'erreur égal à 0,169, et ainsi de suite pour les différents segments terminaux.

Remarque concernant les individus anonymes
Pour les individus anonymes, repérés par une valeur manquante pour la variable de groupe, le programme fournit leur identification et leur classe d'affectation.